Springer Series in
Computational
Mathematics

31

Ernst Hairer
Christian Lubich
Gerhard Wanner

Geometric Numerical Integration

Structure-Preserving Algorithms for Ordinary Differential Equations

Second Edition

With 146 Figures



Ernst Hairer
Gerhard Wanner
Section de Mathématiques
Université de Genève
2-4 rue du Lièvre, C.P. 64
CH-1211 Genève 4, Switzerland
email: Ernst.Hairer@math.unige.ch
Gerhard.Wanner@math.unige.ch

Christian Lubich
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen, Germany
email: Lubich@na.uni-tuebingen.de

Library of Congress Control Number: 2005938386

Mathematics Subject Classification (2000): 65Lxx, 65P10, 70Fxx, 34Cxx

ISSN 0179-3632
ISBN-10 3-540-30663-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-30663-4 Springer Berlin Heidelberg New York
ISBN-10 3-540-43003-2 1st Edition Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Printed in The Netherlands

Typesetting: by the authors and TechBooks using a Springer LaTeX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11592242 46/TechBooks 5 4 3 2 1 0

Preface to the First Edition

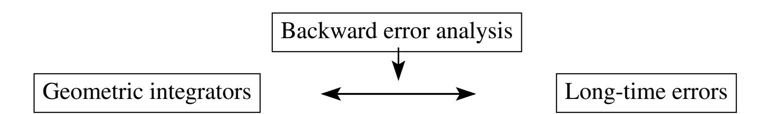
They throw geometry out the door, and it comes back through the window.
(H.G.Forder, Auckland 1973, reading new mathematics at the age of 84)

The subject of this book is numerical methods that preserve geometric properties of the flow of a differential equation: symplectic integrators for Hamiltonian systems, symmetric integrators for reversible systems, methods preserving first integrals and numerical methods on manifolds, including Lie group methods and integrators for constrained mechanical systems, and methods for problems with highly oscillatory solutions. Structure preservation – with its questions as to where, how, and what for – is the unifying theme.

In the last few decades, the theory of numerical methods for general (non-stiff and stiff) ordinary differential equations has reached a certain maturity, and excellent general-purpose codes, mainly based on Runge–Kutta methods or linear multistep methods, have become available. The motivation for developing structure-preserving algorithms for special classes of problems came independently from such different areas of research as astronomy, molecular dynamics, mechanics, theoretical physics, and numerical analysis as well as from other areas of both applied and pure mathematics. It turned out that the preservation of geometric properties of the flow not only produces an improved qualitative behaviour, but also allows for a more accurate long-time integration than with general-purpose methods.

An important shift of view-point came about by ceasing to concentrate on the numerical approximation of a single solution trajectory and instead to consider a numerical method as a *discrete dynamical system* which approximates the flow of the differential equation – and so the geometry of phase space comes back again through the window. This view allows a clear understanding of the preservation of invariants and of methods on manifolds, of symmetry and reversibility of methods, and of the symplecticity of methods and various generalizations. These subjects are presented in Chapters IV through VII of this book. Chapters I through III are of an introductory nature and present examples and numerical integrators together with important parts of the classical order theories and their recent extensions. Chapter VIII deals with questions of numerical implementations and numerical merits of the various methods.

It remains to explain the relationship between geometric properties of the numerical method and the favourable error propagation in long-time integrations. This



is done using the idea of *backward error analysis*, where the numerical one-step map is interpreted as (almost) the flow of a modified differential equation, which is constructed as an asymptotic series (Chapter IX). In this way, geometric properties of the numerical integrator translate into structure preservation on the level of the modified equations. Much insight and rigorous error estimates over long time intervals can then be obtained by combining this backward error analysis with KAM theory and related perturbation theories. This is explained in Chapters X through XII for Hamiltonian and reversible systems. The final Chapters XIII and XIV treat the numerical solution of differential equations with high-frequency oscillations and the long-time dynamics of multistep methods, respectively.

This book grew out of the lecture notes of a course given by Ernst Hairer at the University of Geneva during the academic year 1998/99. These lectures were directed at students in the third and fourth year. The reactions of students as well as of many colleagues, who obtained the notes from the Web, encouraged us to elaborate our ideas to produce the present monograph.

We want to thank all those who have helped and encouraged us to prepare this book. In particular, Martin Hairer for his valuable help in installing computers and his expertise in Latex and Postscript, Jeff Cash and Robert Chan for reading the whole text and correcting countless scientific obscurities and linguistic errors, Haruo Yoshida for making many valuable suggestions, Stéphane Cirilli for preparing the files for all the photographs, and Bernard Dudez, the irreplaceable director of the mathematics library in Geneva. We are also grateful to many friends and colleagues for reading parts of the manuscript and for valuable remarks and discussions, in particular to Assyr Abdulle, Melanie Beck, Sergio Blanes, John Butcher, Mari Paz Calvo, Begoña Cano, Philippe Chartier, David Cohen, Peter Deuflhard, Stig Faltinsen, Francesco Fassò, Martin Gander, Marlis Hochbruck, Bulent Karasözen, Wilhelm Kaup, Ben Leimkuhler, Pierre Leone, Frank Loose, Katina Lorenz, Robert McLachlan, Ander Murua, Alexander Ostermann, Truong Linh Pham, Sebastian Reich, Chus Sanz-Serna, Zaijiu Shang, Yifa Tang, Matt West, Will Wright.

We are especially grateful to Thanh-Ha Le Thi and Dr. Martin Peters from Springer-Verlag Heidelberg for assistance, in particular for their help in getting most of the original photographs from the Oberwolfach Archive and from Springer New York, and for clarifying doubts concerning the copyright.

Geneva and Tübingen, November 2001 The Authors

Preface to the Second Edition

The fast development of the subject – and the fast development of the sales of the first edition of this book – has given the authors the opportunity to prepare this second edition. First of all we have corrected several misprints and minor errors which we have discovered or which have been kindly communicated to us by several readers and colleagues. We cordially thank all of them for their help and for their interest in our work. A major point of confusion has been revealed by Robert McLachlan in his book review in SIAM Reviews.

Besides many details, which have improved the presentation throughout the book, there are the following major additions and changes which make the book about 130 pages longer:

- a more prominent place of the Störmer–Verlet method in the exposition and the examples of the first chapter;
- a discussion of the Hénon–Heiles model as an example of a chaotic Hamiltonian system;
- a new Sect. IV.9 on geometric numerical linear algebra considering differential equations on Stiefel and Grassmann manifolds and dynamical low-rank approximations;
- a new improved composition method of order 10 in Sect. V.3;
- a characterization of B-series methods that conserve quadratic first integrals and a criterion for conjugate symplecticity in Sect. VI.8;
- the section on volume preservation taken from Chap. VII to Chap. VI;
- an extended and more coherent Chap. VII, renamed Non-Canonical Hamiltonian Systems, with more emphasis on the relationships between Hamiltonian systems on manifolds and Poisson systems;
- a completely reorganized and augmented Sect. VII.5 on the rigid body dynamics and Lie–Poisson systems;
- a new Sect. VII.6 on reduced Hamiltonian models of quantum dynamics and Poisson integrators for their numerical treatment;
- an improved step-size control for reversible methods in Sects. VIII.3.2 and IX.6;
- extension of Sect. IX.5 on modified equations of methods on manifolds to include constrained Hamiltonian systems and Lie–Poisson integrators;
- reorganization of Sects. IX.9 and IX.10; study of non-symplectic B-series methods that have a modified Hamiltonian, and counter-examples for symmetric methods showing linear growth in the energy error;

- a more precise discussion of integrable reversible systems with new examples in Chap. XI;
- extension of Chap. XIII on highly oscillatory problems to systems with several constant frequencies and to systems with non-constant mass matrix;
- a new Chap. XIV on oscillatory Hamiltonian systems with time- or solution-dependent high frequencies, emphasizing adiabatic transformations, adiabatic invariants, and adiabatic integrators;
- a completely rewritten Chap. XV with more emphasis on linear multistep methods for second order differential equations; a complete backward error analysis including parasitic modified differential equations; a study of the long-time stability and a rigorous explanation of the long-time near-conservation of energy and angular momentum.

Let us hope that this second revised edition will again meet good acceptance by our readers.

Geneva and Tübingen, October 2005 The Authors

Table of Contents

I. Examples and Numerical Experiments 1
I.1 First Problems and Methods 1
I.1.1 The Lotka–Volterra Model 1
I.1.2 First Numerical Methods 3
I.1.3 The Pendulum as a Hamiltonian System 4
I.1.4 The Störmer–Verlet Scheme 7
I.2 The Kepler Problem and the Outer Solar System 8
I.2.1 Angular Momentum and Kepler's Second Law 9
I.2.2 Exact Integration of the Kepler Problem 10
I.2.3 Numerical Integration of the Kepler Problem 12
I.2.4 The Outer Solar System 13
I.3 The Hénon–Heiles Model 15
I.4 Molecular Dynamics 18
I.5 Highly Oscillatory Problems 21
I.5.1 A Fermi–Pasta–Ulam Problem 21
I.5.2 Application of Classical Integrators 23
I.6 Exercises 24

II. Numerical Integrators 27
II.1 Runge–Kutta and Collocation Methods 27
II.1.1 Runge–Kutta Methods 28
II.1.2 Collocation Methods 30
II.1.3 Gauss and Lobatto Collocation 34
II.1.4 Discontinuous Collocation Methods 35
II.2 Partitioned Runge–Kutta Methods 38
II.2.1 Definition and First Examples 38
II.2.2 Lobatto IIIA–IIIB Pairs 40
II.2.3 Nyström Methods 41
II.3 The Adjoint of a Method 42
II.4 Composition Methods 43
II.5 Splitting Methods 47
II.6 Exercises 50

III. Order Conditions, Trees and B-Series 51
III.1 Runge–Kutta Order Conditions and B-Series 51
III.1.1 Derivation of the Order Conditions 51
III.1.2 B-Series 56
III.1.3 Composition of Methods 59
III.1.4 Composition of B-Series 61
III.1.5 The Butcher Group 64
III.2 Order Conditions for Partitioned Runge–Kutta Methods 66
III.2.1 Bi-Coloured Trees and P-Series 66
III.2.2 Order Conditions for Partitioned Runge–Kutta Methods 68
III.2.3 Order Conditions for Nyström Methods 69
III.3 Order Conditions for Composition Methods 71
III.3.1 Introduction 71
III.3.2 The General Case 73
III.3.3 Reduction of the Order Conditions 75
III.3.4 Order Conditions for Splitting Methods 80
III.4 The Baker-Campbell-Hausdorff Formula 83
III.4.1 Derivative of the Exponential and Its Inverse 83
III.4.2 The BCH Formula 84
III.5 Order Conditions via the BCH Formula 87
III.5.1 Calculus of Lie Derivatives 87
III.5.2 Lie Brackets and Commutativity 89
III.5.3 Splitting Methods 91
III.5.4 Composition Methods 92
III.6 Exercises 95

IV. Conservation of First Integrals and Methods on Manifolds 97
IV.1 Examples of First Integrals 97
IV.2 Quadratic Invariants 101
IV.2.1 Runge–Kutta Methods 101
IV.2.2 Partitioned Runge–Kutta Methods 102
IV.2.3 Nyström Methods 104
IV.3 Polynomial Invariants 105
IV.3.1 The Determinant as a First Integral 105
IV.3.2 Isospectral Flows 107
IV.4 Projection Methods 109
IV.5 Numerical Methods Based on Local Coordinates 113
IV.5.1 Manifolds and the Tangent Space 114
IV.5.2 Differential Equations on Manifolds 115
IV.5.3 Numerical Integrators on Manifolds 116
IV.6 Differential Equations on Lie Groups 118
IV.7 Methods Based on the Magnus Series Expansion 121
IV.8 Lie Group Methods 123
IV.8.1 Crouch-Grossman Methods 124
IV.8.2 Munthe-Kaas Methods 125

IV.8.3 Further Coordinate Mappings 128
IV.9 Geometric Numerical Integration Meets Geometric Numerical Linear Algebra 131
IV.9.1 Numerical Integration on the Stiefel Manifold 131
IV.9.2 Differential Equations on the Grassmann Manifold 135
IV.9.3 Dynamical Low-Rank Approximation 137
IV.10 Exercises 139

V. Symmetric Integration and Reversibility 143
V.1 Reversible Differential Equations and Maps 143
V.2 Symmetric Runge–Kutta Methods 146
V.2.1 Collocation and Runge–Kutta Methods 146
V.2.2 Partitioned Runge–Kutta Methods 148
V.3 Symmetric Composition Methods 149
V.3.1 Symmetric Composition of First Order Methods 150
V.3.2 Symmetric Composition of Symmetric Methods 154
V.3.3 Effective Order and Processing Methods 158
V.4 Symmetric Methods on Manifolds 161
V.4.1 Symmetric Projection 161
V.4.2 Symmetric Methods Based on Local Coordinates 166
V.5 Energy – Momentum Methods and Discrete Gradients 171
V.6 Exercises 176

VI. Symplectic Integration of Hamiltonian Systems 179
VI.1 Hamiltonian Systems 180
VI.1.1 Lagrange's Equations 180
VI.1.2 Hamilton's Canonical Equations 181
VI.2 Symplectic Transformations 182
VI.3 First Examples of Symplectic Integrators 187
VI.4 Symplectic Runge–Kutta Methods 191
VI.4.1 Criterion of Symplecticity 191
VI.4.2 Connection Between Symplectic and Symmetric Methods 194
VI.5 Generating Functions 195
VI.5.1 Existence of Generating Functions 195
VI.5.2 Generating Function for Symplectic Runge–Kutta Methods 198
VI.5.3 The Hamilton–Jacobi Partial Differential Equation 200
VI.5.4 Methods Based on Generating Functions 203
VI.6 Variational Integrators 204
VI.6.1 Hamilton's Principle 204
VI.6.2 Discretization of Hamilton's Principle 206
VI.6.3 Symplectic Partitioned Runge–Kutta Methods Revisited 208
VI.6.4 Noether's Theorem 210

VI.7 Characterization of Symplectic Methods 212
VI.7.1 B-Series Methods Conserving Quadratic First Integrals 212
VI.7.2 Characterization of Symplectic P-Series (and B-Series) 217
VI.7.3 Irreducible Runge–Kutta Methods 220
VI.7.4 Characterization of Irreducible Symplectic Methods . . . 222
VI.8 Conjugate Symplecticity 222
VI.8.1 Examples and Order Conditions 223
VI.8.2 Near Conservation of Quadratic First Integrals 225
VI.9 Volume Preservation 227
VI.10 Exercises 233

VII. Non-Canonical Hamiltonian Systems 237
VII.1 Constrained Mechanical Systems 237
VII.1.1 Introduction and Examples 237
VII.1.2 Hamiltonian Formulation 239
VII.1.3 A Symplectic First Order Method 242
VII.1.4 SHAKE and RATTLE 245
VII.1.5 The Lobatto IIIA - IIIB Pair 247
VII.1.6 Splitting Methods 252
VII.2 Poisson Systems 254
VII.2.1 Canonical Poisson Structure 254
VII.2.2 General Poisson Structures 256
VII.2.3 Hamiltonian Systems on Symplectic Submanifolds 258
VII.3 The Darboux–Lie Theorem 261
VII.3.1 Commutativity of Poisson Flows and Lie Brackets 261
VII.3.2 Simultaneous Linear Partial Differential Equations 262
VII.3.3 Coordinate Changes and the Darboux–Lie Theorem . . . 265
VII.4 Poisson Integrators 268
VII.4.1 Poisson Maps and Symplectic Maps 268
VII.4.2 Poisson Integrators 270
VII.4.3 Integrators Based on the Darboux–Lie Theorem 272
VII.5 Rigid Body Dynamics and Lie–Poisson Systems 274
VII.5.1 History of the Euler Equations 275
VII.5.2 Hamiltonian Formulation of Rigid Body Motion 278
VII.5.3 Rigid Body Integrators 280
VII.5.4 Lie–Poisson Systems 286
VII.5.5 Lie–Poisson Reduction 289
VII.6 Reduced Models of Quantum Dynamics 293
VII.6.1 Hamiltonian Structure of the Schrödinger Equation . . . 293
VII.6.2 The Dirac–Frenkel Variational Principle 295
VII.6.3 Gaussian Wavepacket Dynamics 296
VII.6.4 A Splitting Integrator for Gaussian Wavepackets 298
VII.7 Exercises 301

VIII. Structure-Preserving Implementation 303
VIII.1 Dangers of Using Standard Step Size Control 303
VIII.2 Time Transformations 306
VIII.2.1 Symplectic Integration 306
VIII.2.2 Reversible Integration 309
VIII.3 Structure-Preserving Step Size Control 310
VIII.3.1 Proportional, Reversible Controllers 310
VIII.3.2 Integrating, Reversible Controllers 314
VIII.4 Multiple Time Stepping 316
VIII.4.1 Fast-Slow Splitting: the Impulse Method 317
VIII.4.2 Averaged Forces 319
VIII.5 Reducing Rounding Errors 322
VIII.6 Implementation of Implicit Methods 325
VIII.6.1 Starting Approximations 326
VIII.6.2 Fixed-Point Versus Newton Iteration 330
VIII.7 Exercises 335

IX. Backward Error Analysis and Structure Preservation 337
IX.1 Modified Differential Equation – Examples 337
IX.2 Modified Equations of Symmetric Methods 342
IX.3 Modified Equations of Symplectic Methods 343
IX.3.1 Existence of a Local Modified Hamiltonian 343
IX.3.2 Existence of a Global Modified Hamiltonian 344
IX.3.3 Poisson Integrators 347
IX.4 Modified Equations of Splitting Methods 348
IX.5 Modified Equations of Methods on Manifolds 350
IX.5.1 Methods on Manifolds and First Integrals 350
IX.5.2 Constrained Hamiltonian Systems 352
IX.5.3 Lie–Poisson Integrators 354
IX.6 Modified Equations for Variable Step Sizes 356
IX.7 Rigorous Estimates – Local Error 358
IX.7.1 Estimation of the Derivatives of the Numerical Solution 360
IX.7.2 Estimation of the Coefficients of the Modified Equation 362
IX.7.3 Choice of N and the Estimation of the Local Error 364
IX.8 Long-Time Energy Conservation 366
IX.9 Modified Equation in Terms of Trees 369
IX.9.1 B-Series of the Modified Equation 369
IX.9.2 Elementary Hamiltonians 373
IX.9.3 Modified Hamiltonian 375
IX.9.4 First Integrals Close to the Hamiltonian 375
IX.9.5 Energy Conservation: Examples and Counter-Examples 379
IX.10 Extension to Partitioned Systems 381
IX.10.1 P-Series of the Modified Equation 381
IX.10.2 Elementary Hamiltonians 384
IX.11 Exercises 386

X. Hamiltonian Perturbation Theory and Symplectic Integrators 389
X.1 Completely Integrable Hamiltonian Systems 390
X.1.1 Local Integration by Quadrature 390
X.1.2 Completely Integrable Systems 393
X.1.3 Action-Angle Variables 397
X.1.4 Conditionally Periodic Flows 399
X.1.5 The Toda Lattice – an Integrable System 402
X.2 Transformations in the Perturbation Theory for Integrable Systems 404
X.2.1 The Basic Scheme of Classical Perturbation Theory . . . 405
X.2.2 Lindstedt–Poincaré Series 406
X.2.3 Kolmogorov's Iteration 410
X.2.4 Birkhoff Normalization Near an Invariant Torus 412
X.3 Linear Error Growth and Near-Preservation of First Integrals . . . 413
X.4 Near-Invariant Tori on Exponentially Long Times 417
X.4.1 Estimates of Perturbation Series 417
X.4.2 Near-Invariant Tori of Perturbed Integrable Systems . . . 421
X.4.3 Near-Invariant Tori of Symplectic Integrators 422
X.5 Kolmogorov's Theorem on Invariant Tori 423
X.5.1 Kolmogorov's Theorem 423
X.5.2 KAM Tori under Symplectic Discretization 428
X.6 Invariant Tori of Symplectic Maps 430
X.6.1 A KAM Theorem for Symplectic Near-Identity Maps . 431
X.6.2 Invariant Tori of Symplectic Integrators 433
X.6.3 Strongly Non-Resonant Step Sizes 433
X.7 Exercises 434

XI. Reversible Perturbation Theory and Symmetric Integrators 437
XI.1 Integrable Reversible Systems 437
XI.2 Transformations in Reversible Perturbation Theory 442
XI.2.1 The Basic Scheme of Reversible Perturbation Theory . . 443
XI.2.2 Reversible Perturbation Series 444
XI.2.3 Reversible KAM Theory 445
XI.2.4 Reversible Birkhoff-Type Normalization 447
XI.3 Linear Error Growth and Near-Preservation of First Integrals . . . 448
XI.4 Invariant Tori under Reversible Discretization 451
XI.4.1 Near-Invariant Tori over Exponentially Long Times . . . 451
XI.4.2 A KAM Theorem for Reversible Near-Identity Maps . . 451
XI.5 Exercises 453

XII. Dissipatively Perturbed Hamiltonian and Reversible Systems 455
XII.1 Numerical Experiments with Van der Pol's Equation 455
XII.2 Averaging Transformations 458
XII.2.1 The Basic Scheme of Averaging 458
XII.2.2 Perturbation Series 459

XII.3 Attractive Invariant Manifolds 460
XII.4 Weakly Attractive Invariant Tori of Perturbed Integrable Systems 464
XII.5 Weakly Attractive Invariant Tori of Numerical Integrators 465
XII.5.1 Modified Equations of Perturbed Differential Equations 466
XII.5.2 Symplectic Methods 467
XII.5.3 Symmetric Methods 469
XII.6 Exercises 469

XIII. Oscillatory Differential Equations with Constant High Frequencies . 471
XIII.1 Towards Longer Time Steps in Solving Oscillatory Equations of Motion 471
XIII.1.1 The Störmer–Verlet Method vs. Multiple Time Scales . 472
XIII.1.2 Gautschi's and Deuflhard's Trigonometric Methods . . . 473
XIII.1.3 The Impulse Method 475
XIII.1.4 The Mollified Impulse Method 476
XIII.1.5 Gautschi's Method Revisited 477
XIII.1.6 Two-Force Methods 478
XIII.2 A Nonlinear Model Problem and Numerical Phenomena 478
XIII.2.1 Time Scales in the Fermi–Pasta–Ulam Problem 479
XIII.2.2 Numerical Methods 481
XIII.2.3 Accuracy Comparisons 482
XIII.2.4 Energy Exchange between Stiff Components 483
XIII.2.5 Near-Conservation of Total and Oscillatory Energy 484
XIII.3 Principal Terms of the Modulated Fourier Expansion 486
XIII.3.1 Decomposition of the Exact Solution 486
XIII.3.2 Decomposition of the Numerical Solution 488
XIII.4 Accuracy and Slow Exchange 490
XIII.4.1 Convergence Properties on Bounded Time Intervals . . . 490
XIII.4.2 Intra-Oscillatory and Oscillatory-Smooth Exchanges . . 494
XIII.5 Modulated Fourier Expansions 496
XIII.5.1 Expansion of the Exact Solution 496
XIII.5.2 Expansion of the Numerical Solution 498
XIII.5.3 Expansion of the Velocity Approximation 502
XIII.6 Almost-Invariants of the Modulated Fourier Expansions 503
XIII.6.1 The Hamiltonian of the Modulated Fourier Expansion . 503
XIII.6.2 A Formal Invariant Close to the Oscillatory Energy . . . 505
XIII.6.3 Almost-Invariants of the Numerical Method 507
XIII.7 Long-Time Near-Conservation of Total and Oscillatory Energy . 510
XIII.8 Energy Behaviour of the Störmer–Verlet Method 513
XIII.9 Systems with Several Constant Frequencies 516
XIII.9.1 Oscillatory Energies and Resonances 517
XIII.9.2 Multi-Frequency Modulated Fourier Expansions 519
XIII.9.3 Almost-Invariants of the Modulation System 521
XIII.9.4 Long-Time Near-Conservation of Total and Oscillatory Energies 524

XIII.10 Systems with Non-Constant Mass Matrix . . . 526
XIII.11 Exercises . . . 529

XIV. Oscillatory Differential Equations with Varying High Frequencies . . 531
XIV.1 Linear Systems with Time-Dependent Skew-Hermitian Matrix . . 531
XIV.1.1 Adiabatic Transformation and Adiabatic Invariants 531
XIV.1.2 Adiabatic Integrators . . . 536
XIV.2 Mechanical Systems with Time-Dependent Frequencies . . . 539
XIV.2.1 Canonical Transformation to Adiabatic Variables . . . 540
XIV.2.2 Adiabatic Integrators . . . 547
XIV.2.3 Error Analysis of the Impulse Method . . . 550
XIV.2.4 Error Analysis of the Mollified Impulse Method . . . 554
XIV.3 Mechanical Systems with Solution-Dependent Frequencies . . . 555
XIV.3.1 Constraining Potentials . . . 555
XIV.3.2 Transformation to Adiabatic Variables . . . 558
XIV.3.3 Integrators in Adiabatic Variables . . . 563
XIV.3.4 Analysis of Multiple Time-Stepping Methods . . . 564
XIV.4 Exercises . . . 564

XV. Dynamics of Multistep Methods . . . 567
XV.1 Numerical Methods and Experiments . . . 567
XV.1.1 Linear Multistep Methods . . . 567
XV.1.2 Multistep Methods for Second Order Equations . . . 569
XV.1.3 Partitioned Multistep Methods . . . 572
XV.2 The Underlying One-Step Method . . . 573
XV.2.1 Strictly Stable Multistep methods . . . 573
XV.2.2 Formal Analysis for Weakly Stable Methods . . . 575
XV.3 Backward Error Analysis . . . 576
XV.3.1 Modified Equation for Smooth Numerical Solutions . . . 576
XV.3.2 Parasitic Modified Equations . . . 579
XV.4 Can Multistep Methods be Symplectic? . . . 585
XV.4.1 Non-Symplecticity of the Underlying One-Step Method 585
XV.4.2 Symplecticity in the Higher-Dimensional Phase Space . 587
XV.4.3 Modified Hamiltonian of Multistep Methods . . . 589
XV.4.4 Modified Quadratic First Integrals . . . 591
XV.5 Long-Term Stability . . . 592
XV.5.1 Role of Growth Parameters . . . 592
XV.5.2 Hamiltonian of the Full Modified System . . . 594
XV.5.3 Long-Time Bounds for Parasitic Solution Components 596
XV.6 Explanation of the Long-Time Behaviour . . . 600
XV.6.1 Conservation of Energy and Angular Momentum . . . 600
XV.6.2 Linear Error Growth for Integrable Systems . . . 601
XV.7 Practical Considerations . . . 602
XV.7.1 Numerical Instabilities and Resonances . . . 602
XV.7.2 Extension to Variable Step Sizes . . . 605

XV.8 Multi-Value or General Linear Methods 609
XV.8.1 Underlying One-Step Method and Backward Error Analysis 609
XV.8.2 Symplecticity and Symmetry 611
XV.8.3 Growth Parameters 614
XV.9 Exercises 615

Bibliography 617

Index 637

Chapter I.
Examples and Numerical Experiments

This chapter introduces some interesting examples of differential equations and illustrates different types of qualitative behaviour of numerical methods. We deliberately consider only very simple numerical methods of orders 1 and 2 to emphasize the qualitative aspects of the experiments. The same effects (on a different scale) occur with more sophisticated higher-order integration schemes. The experiments presented here should serve as a motivation for the theoretical and practical investigations of later chapters. The reader is encouraged to repeat the experiments or to invent similar ones.

I.1 First Problems and Methods

> Numerical applications of the case of two dependent variables are not easily obtained. (A.J. Lotka 1925, p. 79)

Our first problems, the Lotka–Volterra model and the pendulum equation, are differential equations in two dimensions and show already many interesting geometric properties. Our first methods are various variants of the Euler method, the midpoint rule, and the Störmer–Verlet scheme.

I.1.1 The Lotka–Volterra Model

We start with an equation from mathematical biology which models the growth of animal species. If a real variable $u(t)$ is to represent the number of individuals of a certain species at time t, the simplest assumption about its evolution is $du/dt = u{\cdot}\alpha$, where α is the reproduction rate. A constant α leads to exponential growth. In the case of more species living together, the reproduction rates will also depend on the population numbers of the *other* species. For example, for two species with $u(t)$ denoting the number of predators and $v(t)$ the number of prey, a plausible assumption is made by the *Lotka–Volterra model*

$$\begin{aligned} \dot u &= u(v-2) \\ \dot v &= v(1-u), \end{aligned} \tag{1.1}$$

where the dots on u and v stand for differentiation with respect to time. (We have chosen the constants 2 and 1 in (1.1) arbitrarily.) A.J. Lotka (1925, Chap. VIII) used

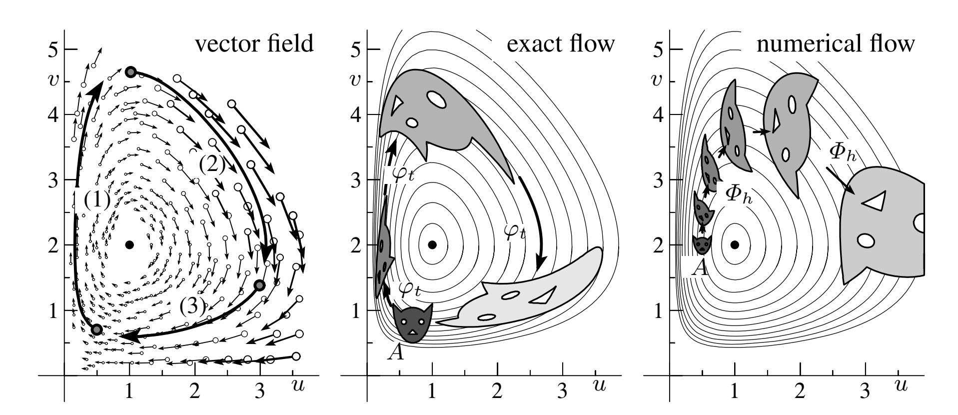


Fig. 1.1. Vector field, exact flow, and numerical flow for the Lotka–Volterra model (1.1)

this model to study parasitic invasion of insect species, and, with its help, V. Volterra (1927) explained curious fishing data from the upper Adriatic Sea following World War I.

Equations (1.1) constitute an autonomous system of differential equations. In general, we write such a system in the form

$$\dot{y} = f(y) \,. \tag{1.2}$$

Every y represents a point in the *phase space*, in equation (1.1) above $y = (u, v)$ is in the phase plane $\mathbb{R}^2$. The vector-valued function $f(y)$ represents a *vector field* which, at any point of the phase space, prescribes the velocity (direction and speed) of the solution $y(t)$ that passes through that point (see the first picture of Fig. 1.1).

For the Lotka–Volterra model, we observe that the system cycles through three stages: (1) the prey population increases; (2) the predator population increases by feeding on the prey; (3) the predator population diminishes due to lack of food.

Flow of the System. A fundamental concept is the *flow* over time t. This is the mapping which, to any point y_0 in the phase space, associates the value $y(t)$ of the solution with initial value $y(0) = y_0$. This map, denoted by φ_t, is thus defined by

$$\varphi_t(y_0) = y(t) \quad \text{if} \quad y(0) = y_0. \tag{1.3}$$

The second picture of Fig. 1.1 shows the results of three iterations of φ_t (with $t = 1.3$) for the Lotka–Volterra problem, for a set of initial values $y_0 = (u_0, v_0)$ forming an animal-shaped set A.[1]

Invariants. If we divide the two equations of (1.1) by each other, we obtain a single equation between the variables u and v. After separation of variables we get

$$0 = \frac{1-u}{u}\,\dot{u} - \frac{v-2}{v}\,\dot{v} = \frac{d}{dt}\,I(u,v)$$

[1] This cat came to fame through Arnold (1963).

where

$$I(u,v) = \ln u - u + 2\ln v - v, \tag{1.4}$$

so that $I(u(t), v(t)) = Const$ for all t. We call the function I an *invariant* of the system (1.1). Every solution of (1.1) thus lies on a level curve of (1.4). Some of these curves are drawn in the pictures of Fig. 1.1. Since the level curves are closed, all solutions of (1.1) are periodic.

I.1.2 First Numerical Methods

Explicit Euler Method. The simplest of all numerical methods for the system (1.2) is the method formulated by Euler (1768),

$$y_{n+1} = y_n + hf(y_n). \tag{1.5}$$

It uses a constant step size h to compute, one after the other, approximations y_1, y_2, y_3, ... to the values $y(h)$, $y(2h)$, $y(3h)$, ... of the solution starting from a given initial value $y(0) = y_0$. The method is called the *explicit Euler method*, because the approximation y_{n+1} is computed using an explicit evaluation of f at the already known value y_n. Such a formula represents a mapping

$$\Phi_h : y_n \mapsto y_{n+1},$$

which we call the *discrete* or *numerical flow*. Some iterations of the discrete flow for the Lotka–Volterra problem (1.1) (with $h = 0.5$) are represented in the third picture of Fig. 1.1.

Implicit Euler Method. The *implicit Euler method*

$$y_{n+1} = y_n + hf(y_{n+1}), \tag{1.6}$$

is known for its all-damping stability properties. In contrast to (1.5), the approximation y_{n+1} is defined implicitly by (1.6), and the implementation requires the numerical solution of a nonlinear system of equations.

Implicit Midpoint Rule. Taking the mean of y_n and y_{n+1} in the argument of f, we get the *implicit midpoint rule*

$$y_{n+1} = y_n + hf\Big(\frac{y_n + y_{n+1}}{2}\Big). \tag{1.7}$$

It is a *symmetric* method, which means that the formula is left unaltered after exchanging $y_n \leftrightarrow y_{n+1}$ and $h \leftrightarrow -h$ (more on symmetric methods in Chap. V).

Symplectic Euler Methods. For *partitioned* systems

$$\begin{aligned} \dot u &= a(u,v) \\ \dot v &= b(u,v), \end{aligned} \tag{1.8}$$

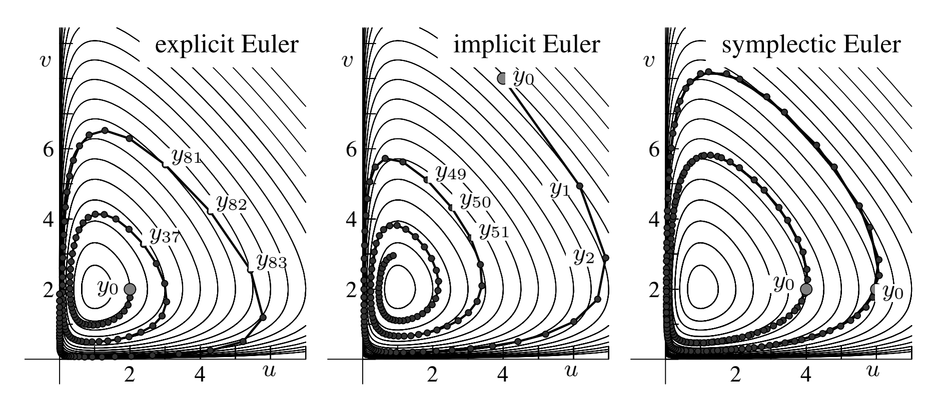


Fig. 1.2. Solutions of the Lotka–Volterra equations (1.1) (step sizes $h = 0.12$; initial values $(2, 2)$ for the explicit Euler method, $(4, 8)$ for the implicit Euler method, $(4, 2)$ and $(6, 2)$ for the symplectic Euler method)

such as the problem (1.1), we consider also *partitioned* Euler methods

$$\begin{aligned} u_{n+1} &= u_n + ha(u_n, v_{n+1}) \\ v_{n+1} &= v_n + hb(u_n, v_{n+1}), \end{aligned} \qquad \text{or} \qquad \begin{aligned} u_{n+1} &= u_n + ha(u_{n+1}, v_n) \\ v_{n+1} &= v_n + hb(u_{n+1}, v_n), \end{aligned} \tag{1.9}$$

which treat one variable by the implicit and the other variable by the explicit Euler method. In view of an important property of this method, discovered by de Vogelaere (1956) and to be discussed in Chap. VI, we call them *symplectic Euler methods*.

Numerical Example for the Lotka–Volterra Problem. Our first numerical experiment shows the behaviour of the various numerical methods applied to the Lotka–Volterra problem. In particular, we are interested in the preservation of the invariant I over long times. Fig. 1.2 plots the numerical approximations of the first 125 steps with the above numerical methods applied to (1.1), all with constant step sizes. We observe that the explicit and implicit Euler methods show wrong qualitative behaviour. The numerical solution either spirals outwards or inwards. The symplectic Euler method (implicit in u and explicit in v), however, gives a numerical solution that lies apparently on a closed curve as does the exact solution. Note that the curves of the numerical and exact solutions do not coincide.

I.1.3 The Pendulum as a Hamiltonian System

A great deal of attention in this book will be addressed to Hamiltonian problems, and our next examples will be of this type. These problems are of the form

$$\dot{p} = -H_q(p, q), \qquad \dot{q} = H_p(p, q), \tag{1.10}$$

where the *Hamiltonian* $H(p_1, \ldots, p_d, q_1, \ldots q_d)$ represents the total energy; q_i are the position coordinates and p_i the momenta for $i = 1, \ldots, d$, with d the number of

degrees of freedom; H_p and H_q are the vectors of partial derivatives. One verifies easily by differentiation (see Sect. IV.1) that, along the solution curves of (1.10),

$$H\big(p(t), q(t)\big) = Const, \tag{1.11}$$

i.e., the Hamiltonian is an invariant or a *first integral*. More details about Hamiltonian systems and their derivation from Lagrangian mechanics will be given in Sect. VI.1.

Pendulum. The mathematical pendulum (mass $m = 1$, massless rod of length $\ell = 1$, gravitational acceleration $g = 1$) is a system with one degree of freedom having the Hamiltonian

$$H(p, q) = \frac{1}{2}\,p^2 - \cos q, \tag{1.12}$$

so that the equations of motion (1.10) become

$$\dot p = -\sin q, \qquad \dot q = p. \tag{1.13}$$

Since the vector field (1.13) is 2π-periodic in q, it is natural to consider q as a variable on the circle S^1. Hence, the phase space of points (p, q) becomes the cylinder $\mathbb{R} \times S^1$. Fig. 1.3 shows some level curves of $H(p, q)$. By (1.11), the solution curves of the problem (1.13) lie on such level curves.

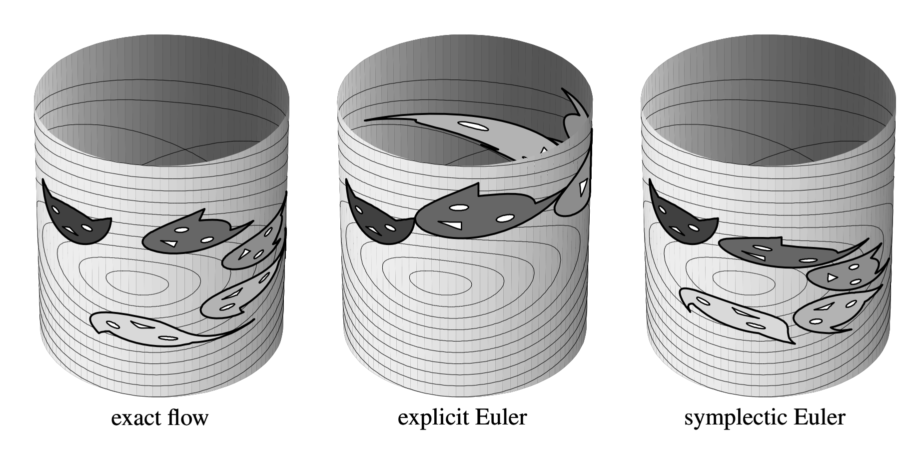


Fig. 1.3. Exact and numerical flow for the pendulum problem (1.13); step sizes $h = t = 1$

Area Preservation. Figure 1.3 (first picture) illustrates that the exact flow of a Hamiltonian system (1.10) is area preserving. This can be explained as follows: the derivative of the flow φ_t with respect to initial values (p, q),

$$\varphi_t'(p, q) = \frac{\partial\big(p(t), q(t)\big)}{\partial(p, q)},$$

satisfies the variational equation [2]

$$\dot{\varphi}_t'(p,q) = \begin{pmatrix} -H_{pq} & -H_{qq} \\ H_{pp} & H_{qp} \end{pmatrix} \varphi_t'(p,q) ,$$

where the second partial derivatives of H are evaluated at $\varphi_t(p,q)$. In the case of one degree of freedom ($d = 1$), a simple computation shows that

$$\frac{d}{dt} \det \varphi_t'(p,q) = \frac{d}{dt}\Big(\frac{\partial p(t)}{\partial p} \frac{\partial q(t)}{\partial q} - \frac{\partial p(t)}{\partial q} \frac{\partial q(t)}{\partial p} \Big) = \ldots = 0.$$

Since φ_0 is the identity, this implies $\det \varphi_t'(p,q) = 1$ for all t, which means that the flow $\varphi_t(p,q)$ is an *area-preserving* mapping.

The last two pictures of Fig. 1.3 show numerical flows. The explicit Euler method is clearly seen not to preserve area but the symplectic Euler method is (this will be proved in Sect. VI.3). One of the aims of 'geometric integration' is the study of numerical integrators that preserve such types of qualitative behaviour of the exact flow.

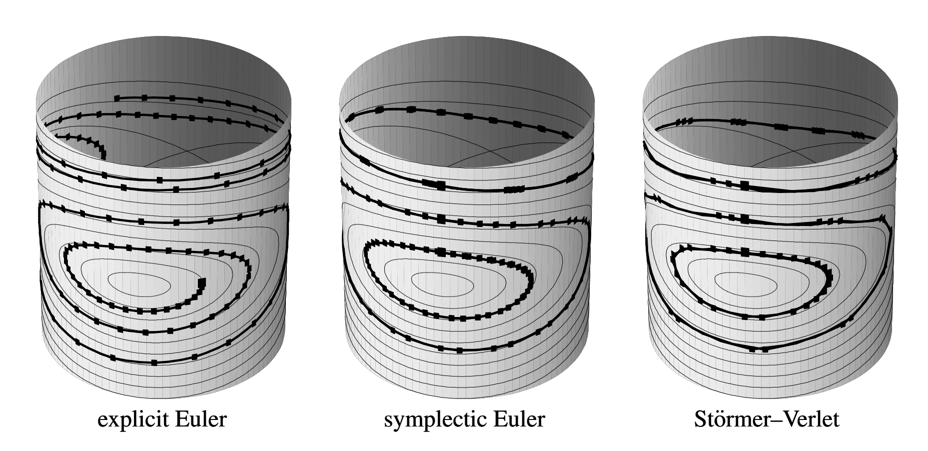


Fig. 1.4. Solutions of the pendulum problem (1.13); explicit Euler with step size $h = 0.2$, initial value $(p_0, q_0) = (0, 0.5)$; symplectic Euler with $h = 0.3$ and initial values $q_0 = 0$, $p_0 = 0.7, 1.4, 2.1$; Störmer–Verlet with $h = 0.6$

Numerical Experiment. We apply the above numerical methods to the pendulum equations (see Fig. 1.4). Similar to the computations for the Lotka–Volterra equations, we observe that the numerical solutions of the explicit Euler and of the implicit Euler method (not drawn in Fig. 1.4) spiral either outwards or inwards. The symplectic Euler method shows the correct qualitative behaviour, but destroys the left-right symmetry of the problem. The Störmer–Verlet scheme, which we discuss next, works perfectly even with doubled step size.

[2] As is common in the study of mechanical problems, we use *dots* for denoting time-derivatives, and we use *primes* for denoting derivatives with respect to other variables.



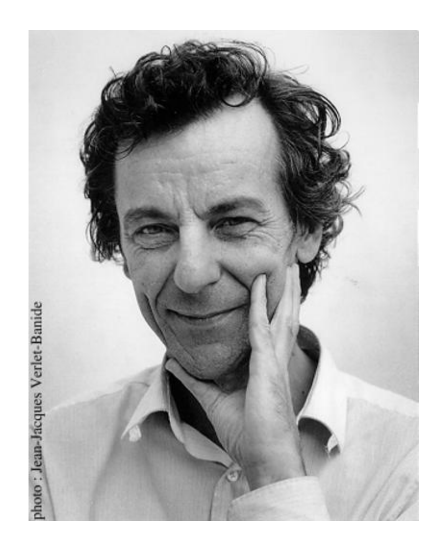


Fig. 1.5. Carl Störmer (left picture), born: 3 September 1874 in Skien (Norway), died: 13 August 1957.
Loup Verlet (right picture), born: 24 May 1931 in Paris

I.1.4 The Störmer–Verlet Scheme

The above equations (1.13) for the pendulum are of the form

$$\begin{aligned} \dot p &= f(q) \\ \dot q &= p \end{aligned} \qquad \text{or} \qquad \ddot q = f(q) \tag{1.14}$$

which is the important special case of a second order differential equation. The most natural discretization of (1.14) is

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f(q_n), \tag{1.15}$$

which is just obtained by replacing the second derivative in (1.14) by the central second-order difference quotient. This basic method, or its equivalent formulation given below, is called the *Störmer method* in astronomy, the *Verlet method* [3] in molecular dynamics, the *leap-frog method* in the context of partial differential equations, and it has further names in other areas (see Hairer, Lubich & Wanner (2003), p. 402). C. Störmer (1907) used higher-order variants for numerical computations concerning the aurora borealis. L. Verlet (1967) proposed this method for computations in molecular dynamics, where it has become by far the most widely used integration scheme.

Geometrically, the Störmer–Verlet method can be seen as produced by parabolas, which in the points t_n possess the right second derivative $f(q_n)$ (see Fig. 1.6

[3] Irony of fate: Professor Loup Verlet, who later became interested in the history of science, discovered precisely “his” method in Newton’s *Principia* (Book I, figure for Theorem I, see Sect. I.2.1 below).

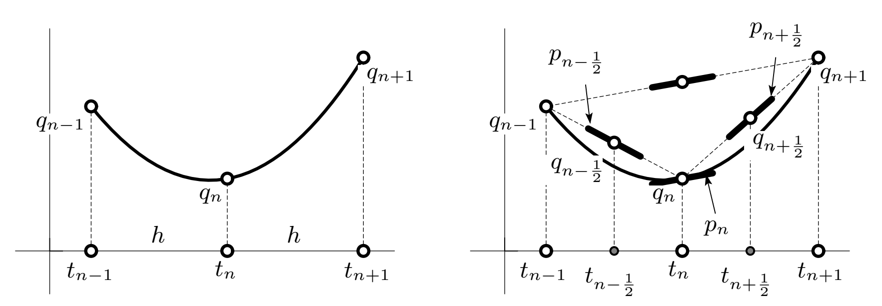


Fig. 1.6. Illustration for the Störmer–Verlet method

to the left). But we can also think of polygons, which possess the right slope in the midpoints (Fig. 1.6 to the right).

Approximations to the derivative $p = \dot{q}$ are simply obtained by

$$p_n = \frac{q_{n+1} - q_{n-1}}{2h} \qquad \text{and} \qquad p_{n+1/2} = \frac{q_{n+1} - q_n}{h}. \tag{1.16}$$

One-Step Formulation. The Störmer–Verlet method admits a one-step formulation which is useful for actual computations. The value q_n together with the slope p_n and the second derivative $f(q_n)$, all at t_n, uniquely determine the parabola and hence also the approximation (p_{n+1}, q_{n+1}) at t_{n+1}. Writing (1.15) as $p_{n+1/2} - p_{n-1/2} = hf(q_n)$ and using $p_{n+1/2} + p_{n-1/2} = 2p_n$, we get by elimination of either $p_{n+1/2}$ or $p_{n-1/2}$ the formulae

$$\begin{aligned} p_{n+1/2} &= p_n + \frac{h}{2} f(q_n) \\ q_{n+1} &= q_n + h p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} + \frac{h}{2} f(q_{n+1}) \end{aligned} \tag{1.17}$$

which is an explicit one-step method $\Phi_h : (q_n, p_n) \mapsto (q_{n+1}, p_{n+1})$ for the corresponding first order system of (1.14). If one is not interested in the values p_n of the derivative, the first and third equations in (1.17) can be replaced by

$$p_{n+1/2} = p_{n-1/2} + h\, f(q_n).$$

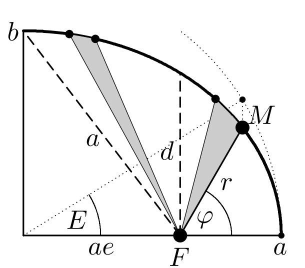
I.2 The Kepler Problem and the Outer Solar System

I awoke as if from sleep, a new light broke on me. (J. Kepler; quoted from J.L.E. Dreyer, *A history of astronomy*, 1906, Dover 1953, p. 391)

One of the great achievements in the history of science was the discovery of the laws of J. Kepler (1609), based on many precise measurements of the positions of Mars by Tycho Brahe and himself. The planets move in *elliptic orbits* with the sun at one of the foci (Kepler's first law)

$$r = \frac{d}{1 + e\cos\varphi} = a - ae\cos E, \tag{2.1}$$

(where a = great axis, e = eccentricity, $b = a\sqrt{1-e^2}$, $d = b\sqrt{1-e^2} = a(1-e^2)$, E = eccentric anomaly, φ = true anomaly).



Newton (*Principia* 1687) then *explained* this motion by his general law of gravitational attraction (proportional to $1/r^2$) and the relation between forces and acceleration (the "Lex II" of the *Principia*). This then opened the way for treating arbitrary celestial motions by solving differential equations.

Two-Body Problem. For computing the motion of two bodies which attract each other, we choose one of the bodies as the centre of our coordinate system; the motion will then stay in a plane (Exercise 3) and we can use two-dimensional coordinates $q = (q_1, q_2)$ for the position of the second body. Newton's laws, with a suitable normalization, then yield the following differential equations

$$\ddot{q}_1 = -\frac{q_1}{(q_1^2 + q_2^2)^{3/2}}, \qquad \ddot{q}_2 = -\frac{q_2}{(q_1^2 + q_2^2)^{3/2}}. \tag{2.2}$$

This is equivalent to a Hamiltonian system with the Hamiltonian

$$H(p_1, p_2, q_1, q_2) = \frac{1}{2}\left(p_1^2 + p_2^2\right) - \frac{1}{\sqrt{q_1^2 + q_2^2}}, \qquad p_i = \dot{q}_i. \tag{2.3}$$

I.2.1 Angular Momentum and Kepler's Second Law

The system has not only the total energy $H(p, q)$ as a first integral, but also the angular momentum

$$L(p_1, p_2, q_1, q_2) = q_1 p_2 - q_2 p_1. \tag{2.4}$$

This can be checked by differentiation and is nothing other than *Kepler's second law*, which says that the ray FM sweeps equal areas in equal times (see the little picture at the beginning of Sect. I.2).

A beautiful *geometric* justification of this law is due to I. Newton[4] (*Principia* (1687), Book I, figure for Theorem I). The idea is to apply the Störmer–Verlet scheme (1.15) to the equations (2.2) (see Fig. 2.1). By hypothesis, the diagonal of the parallelogram $q_{n-1}q_nq_{n+1}$, which is $(q_{n+1} - q_n) - (q_n - q_{n-1}) = q_{n+1} - 2q_n + q_{n-1} = Const \cdot f(q_n)$, points towards the sun S. Therefore, the altitudes of the triangles $q_{n-1}q_nS$ and $q_{n+1}q_nS$ are equal. Since they have the common base q_nS, they also have equal areas. Hence

$$\det(q_{n-1}, q_n - q_{n-1}) = \det(q_n, q_{n+1} - q_n)$$

and by passing to the limit $h \to 0$ we see that $\det(q, p) = Const$. This is (2.4).

[4] We are grateful to a private communication of L. Verlet for this reference

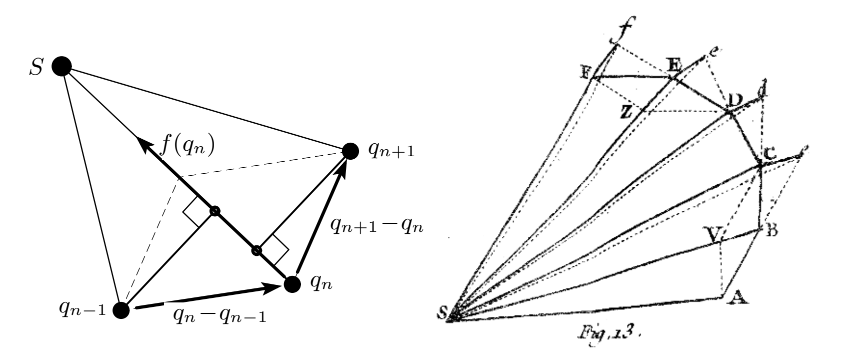


Fig. 2.1. Proof of Kepler's Second Law (left); facsimile from Newton's *Principia* (right)

We have not only an elegant proof for this invariant, but we also see that *the Störmer–Verlet scheme preserves this invariant for every* $h > 0$.

I.2.2 Exact Integration of the Kepler Problem

> Pour voir présentement que cette courbe ABC ... est toûjours une Section Conique, ainsi que Mr. Newton l'a supposé, *pag. 55. Coroll.I.* sans le démontrer; il y faut bien plus d'adresse: (Joh. Bernoulli 1710, p. 475)

It is now interesting, inversely to the procedure of Newton, to prove that *any* solution of (2.2) follows either an elliptic, parabolic or hyperbolic arc and to describe the solutions analytically. This was first done by Joh. Bernoulli (1710, full of sarcasm against Newton), and by Newton (1713, second edition of the *Principia*, without mentioning a word about Bernoulli).

By (2.3) and (2.4), every solution of (2.2) satisfies the two relations

$$\frac{1}{2}\left(\dot q_1^2+\dot q_2^2\right)-\frac{1}{\sqrt{q_1^2+q_2^2}}=H_0, \qquad q_1\dot q_2-q_2\dot q_1=L_0, \tag{2.5}$$

where the constants H_0 and L_0 are determined by the initial values. Using polar coordinates $q_1 = r\cos\varphi$, $q_2 = r\sin\varphi$, this system becomes

$$\frac{1}{2}\left(\dot r^2+r^2\dot\varphi^2\right)-\frac{1}{r}=H_0, \qquad r^2\dot\varphi=L_0. \tag{2.6}$$

For its solution we consider r as a function of φ and write $\dot r = \frac{dr}{d\varphi}\cdot\dot\varphi$. The elimination of $\dot\varphi$ in (2.6) then yields

$$\frac{1}{2}\left(\left(\frac{dr}{d\varphi}\right)^2+r^2\right)\frac{L_0^2}{r^4}-\frac{1}{r}=H_0.$$

In this equation we use the substitution $r = 1/u$, $dr = -du/u^2$, which gives (with $' = d/d\varphi$)

$$\frac{1}{2}(u'^2+u^2)-\frac{u}{L_0^2}-\frac{H_0}{L_0^2}=0. \tag{2.7}$$

This is a "Hamiltonian" for the system

$$u'' + u = \frac{1}{d} \quad \text{i.e.,} \quad u = \frac{1}{d} + c_1 \cos\varphi + c_2 \sin\varphi = \frac{1 + e\cos(\varphi - \varphi^*)}{d} \tag{2.8}$$

where $d = L_0^2$ and the constant e becomes, from (2.7),

$$e^2 = 1 + 2H_0 L_0^2 \tag{2.9}$$

(by Exercise 7, the expression $1+2H_0L_0^2$ is non-negative). This is precisely formula (2.1). The angle φ^* is determined by the initial values r_0 and φ_0. Equation (2.1) represents an elliptic orbit with eccentricity e for $H_0 < 0$ (see Fig. 2.2, dotted line), a parabola for $H_0 = 0$, and a hyperbola for $H_0 > 0$.

Finally, we must determine the variables r and φ as functions of t. With the relation (2.8) and $r = 1/u$, the second equation of (2.6) gives

$$\frac{d^2}{\big(1 + e\cos(\varphi - \varphi^*)\big)^2}\, d\varphi = L_0\, dt \tag{2.10}$$

which, after an elementary, but not easy, integration, represents an implicit equation for $\varphi(t)$.

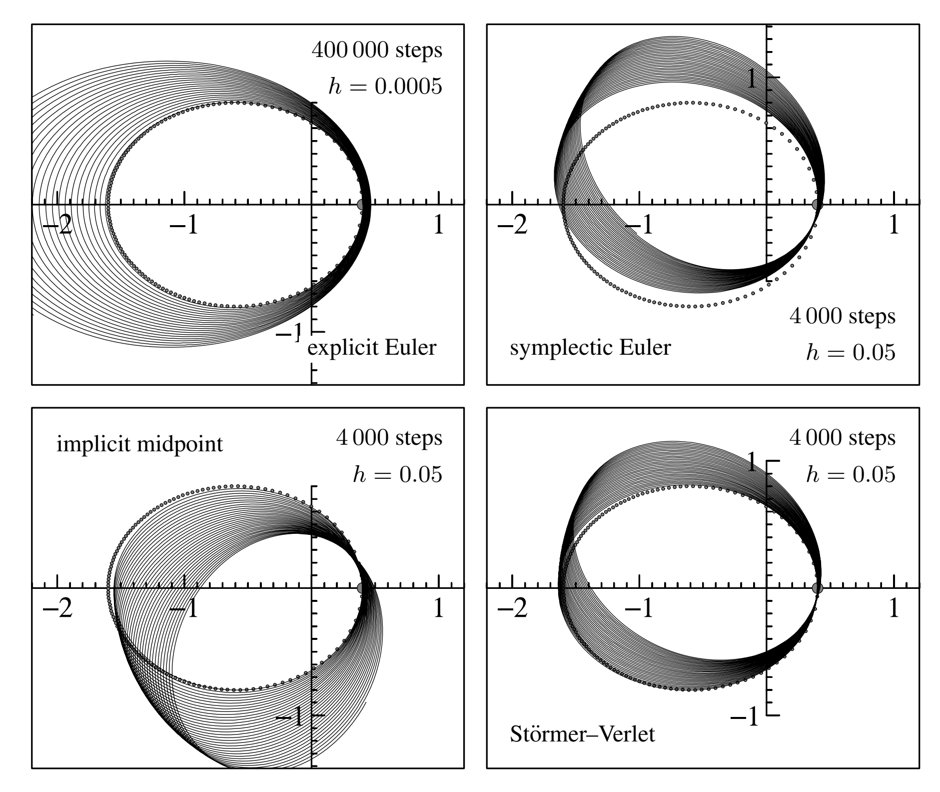


Fig. 2.2. Numerical solutions of the Kepler problem (eccentricity $e = 0.6$; in dots: exact solution)

I.2.3 Numerical Integration of the Kepler Problem

For the problem (2.2) we choose, with $0 \le e < 1$, the initial values

$$q_1(0) = 1-e, \quad q_2(0) = 0, \quad \dot q_1(0) = 0, \quad \dot q_2(0) = \sqrt{\frac{1+e}{1-e}}. \tag{2.11}$$

This implies that $H_0 = -1/2$, $L_0 = \sqrt{1-e^2}$, $d = 1-e^2$ and $\varphi^* = 0$. The period of the solution is 2π (Exercise 5). Fig. 2.2 shows some numerical solutions for the eccentricity $e = 0.6$ compared to the exact solution. After our previous experience, it is no longer a surprise that the explicit Euler method spirals outwards and gives a completely wrong answer. For the other methods we take a step size 100 times larger in order to “see something”. We see that the nonsymmetric symplectic Euler method distorts the ellipse, and that all methods exhibit a *precession* effect, clockwise for Störmer–Verlet and symplectic Euler, anti-clockwise for the implicit midpoint rule. The same behaviour occurs for the exact solution of *perturbed* Kepler problems (Exercise 12) and has occupied astronomers for centuries.

Our next experiment (Fig. 2.3) studies the conservation of invariants and the global error. The main observation is that the error in the energy grows linearly for the explicit Euler method, and it remains bounded and small (no secular terms) for the symplectic Euler method. The global error, measured in the Euclidean norm, shows a quadratic growth for the explicit Euler compared to a linear growth for the symplectic Euler. As indicated in Table 2.1 the implicit midpoint rule and the Störmer–Verlet scheme behave similar to the symplectic Euler, but have a smaller

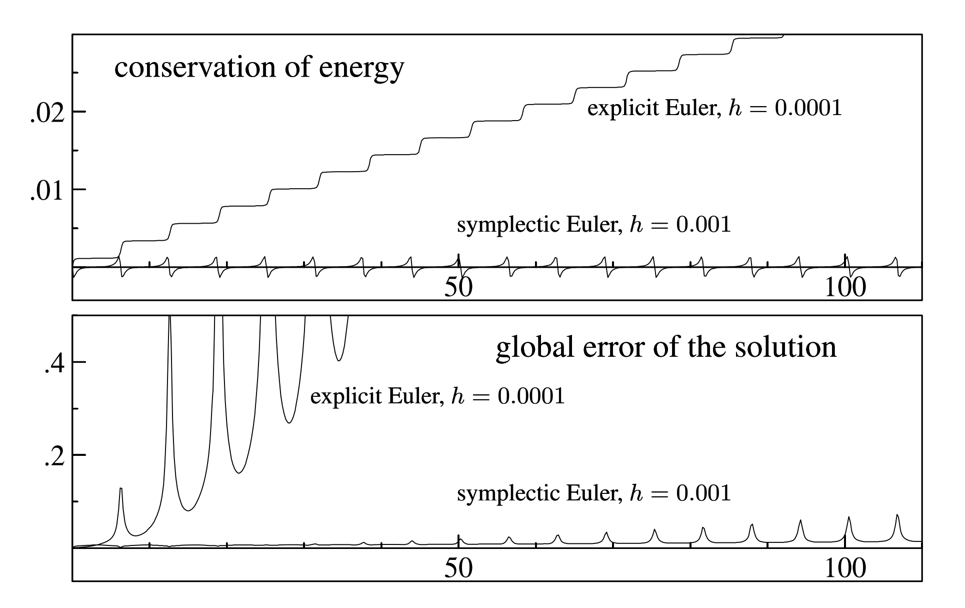


Fig. 2.3. Energy conservation and global error for the Kepler problem

Table 2.1. Qualitative long-time behaviour for the Kepler problem; t is time, h the step size

method	error in H	error in L	global error
explicit Euler	$\mathcal{O}(th)$	$\mathcal{O}(th)$	$\mathcal{O}(t^2h)$
symplectic Euler	$\mathcal{O}(h)$	0	$\mathcal{O}(th)$
implicit midpoint	$\mathcal{O}(h^2)$	0	$\mathcal{O}(th^2)$
Störmer–Verlet	$\mathcal{O}(h^2)$	0	$\mathcal{O}(th^2)$

error due to their higher order. We remark that the angular momentum $L(p,q)$ is exactly conserved by the symplectic Euler, the Störmer–Verlet, and the implicit midpoint rule.

I.2.4 The Outer Solar System

> The evolution of the entire planetary system has been numerically integrated for a time span of nearly 100 million years[5]. This calculation confirms that the evolution of the solar system as a whole is chaotic, . . .
> (G.J. Sussman & J. Wisdom 1992)

We next apply our methods to the system which describes the motion of the five outer planets relative to the sun. This system has been studied extensively by astronomers. The problem is a Hamiltonian system (1.10) (N-body problem) with

$$H(p,q) = \frac{1}{2}\sum_{i=0}^{5}\frac{1}{m_i}\,p_i^T p_i - G\sum_{i=1}^{5}\sum_{j=0}^{i-1}\frac{m_i\,m_j}{\|q_i - q_j\|}. \tag{2.12}$$

Here p and q are the supervectors composed by the vectors $p_i, q_i \in \mathbb{R}^3$ (momenta and positions), respectively. The chosen units are: masses relative to the sun, so that the sun has mass 1. We have taken

$$m_0 = 1.00000597682$$

to take account of the inner planets. Distances are in astronomical units (1 [A.U.] = 149 597 870 [km]), times in earth days, and the gravitational constant is

$$G = 2.95912208286 \cdot 10^{-4}.$$

The initial values for the sun are taken as $q_0(0) = (0,0,0)^T$ and $\dot q_0(0) = (0,0,0)^T$. All other data (masses of the planets and the initial positions and initial velocities) are given in Table 2.2. The initial data is taken from “Ahnerts Kalender für Sternfreunde 1994”, Johann Ambrosius Barth Verlag 1993, and they correspond to September 5, 1994 at 0h00.[6]

[5] 100 million years is not much in astronomical time scales; it just goes back to “Jurassic Park”.

[6] We thank Alexander Ostermann, who provided us with this data.

Table 2.2. Data for the outer solar system

planet	mass	initial position	initial velocity
Jupiter	$m_1 = 0.000954786104043$	-3.5023653 -3.8169847 -1.5507963	0.00565429 -0.00412490 -0.00190589
Saturn	$m_2 = 0.000285583733151$	9.0755314 -3.0458353 -1.6483708	0.00168318 0.00483525 0.00192462
Uranus	$m_3 = 0.0000437273164546$	8.3101420 -16.2901086 -7.2521278	0.00354178 0.00137102 0.00055029
Neptune	$m_4 = 0.0000517759138449$	11.4707666 -25.7294829 -10.8169456	0.00288930 0.00114527 0.00039677
Pluto	$m_5 = 1/(1.3 \cdot 10^8)$	-15.5387357 -25.2225594 -3.1902382	0.00276725 -0.00170702 -0.00136504

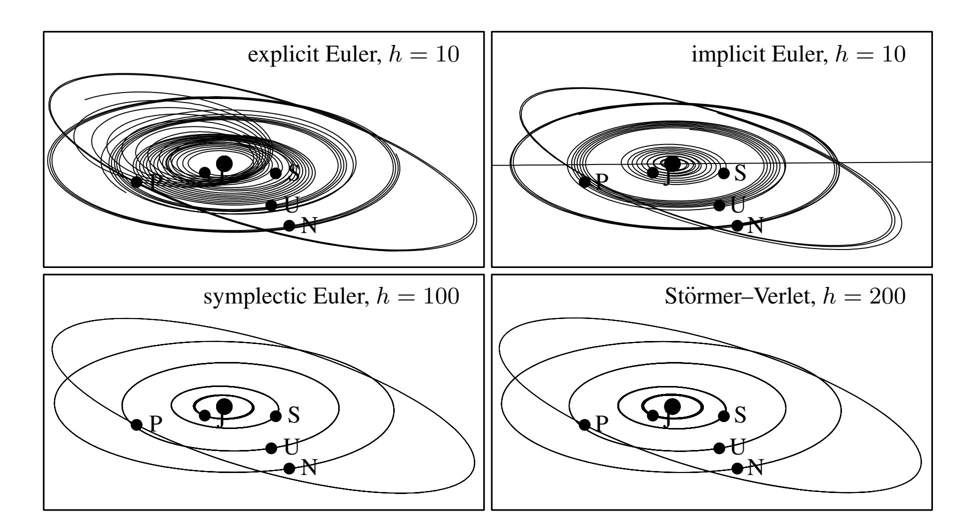


Fig. 2.4. Solutions of the outer solar system

To this system we apply the explicit and implicit Euler methods with step size $h = 10$, the symplectic Euler and the Störmer–Verlet method with much larger step sizes $h = 100$ and $h = 200$, repectively, all over a time period of 200 000 days. The numerical solution (see Fig. 2.4) behaves similarly to that for the Kepler problem. With the explicit Euler method the planets have increasing energy, they spiral outwards, Jupiter approaches Saturn which leaves the plane of the two-body motion. With the implicit Euler method the planets (first Jupiter and then Saturn)

fall into the sun and are thrown far away. Both the symplectic Euler method and the Störmer–Verlet scheme show the correct behaviour. An integration over a much longer time of say several million years does not deteriorate this behaviour. Let us remark that Sussman & Wisdom (1992) have integrated the outer solar system with special geometric integrators.

I.3 The Hénon–Heiles Model

> . . . because: (1) it is analytically simple; this makes the computation of the trajectories easy; (2) at the same time, it is sufficiently complicated to give trajectories which are far from trivial . (Hénon & Heiles 1964)

The Hénon–Heiles model was created for describing stellar motion, followed for a very long time, inside the gravitational potential $U_0(r,z)$ of a galaxy with cylindrical symmetry (Hénon & Heiles 1964). Extensive numerical experimentations should help to answer the question, if there exists, besides the known invariants H and L, a *third* invariant. Despite endless tentatives of analytical calculations during many decades, such a formula had not been found.

After a reduction of the dimension, a Hamiltonian in two degrees of freedom of the form

$$H(p,q) = \frac{1}{2}(p_1^2 + p_2^2) + U(q) \tag{3.1}$$

is obtained and the question is, if such an equation has a *second* invariant. Here, Hénon and Heiles put aside the astronomical origin of the problem and choose

$$U(q) = \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3}q_2^3 \tag{3.2}$$

(see citation). The potential U is represented in Fig. 3.1. When U approaches $\frac{1}{6}$, the level curves of U tend to an equilateral triangle, whose vertices are saddle points of U. The corresponding system

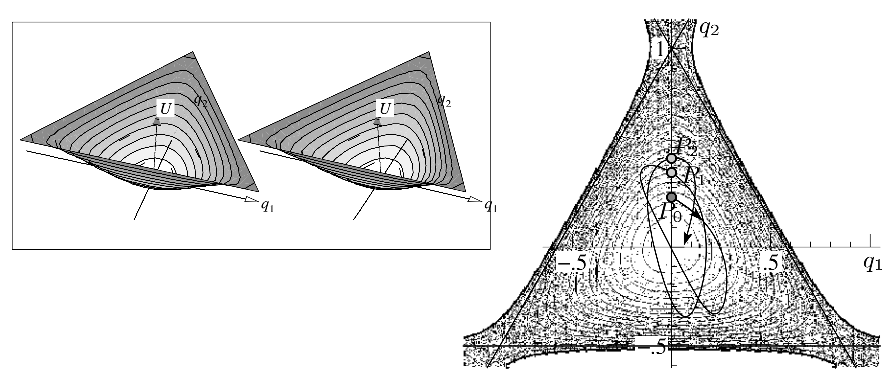


Fig. 3.1. Potential of the Hénon–Heiles Model and a solution

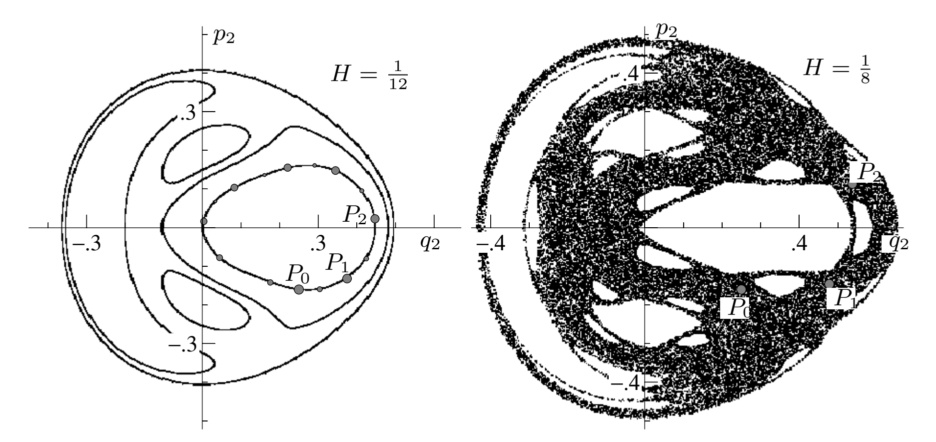


Fig. 3.2. Poincaré cuts for $q_1 = 0, p_1 > 0$ of the Hénon–Heiles Model for $H = \frac{1}{12}$ (6 orbits, left) and $H = \frac{1}{8}$ (1 orbit, right)

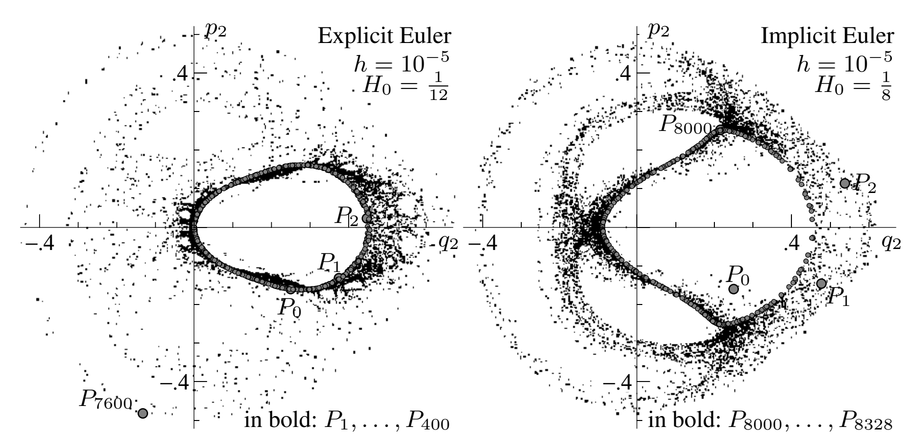


Fig. 3.3. Poincaré cuts for numerical methods, one orbit each; explicit Euler (left), implicit Euler (right). Same initial data as in Fig. 3.2

$$\ddot{q}_1 = -q_1 - 2q_1q_2\,, \qquad \ddot{q}_2 = -q_2 - q_1^2 + q_2^2 \tag{3.3}$$

has solutions with nontrivial properties. For given initial values with $H(p_0, q_0) < \frac{1}{6}$ and q_0 inside the triangle $U \le \frac{1}{6}$, the solution stays there and moves somehow like a mass point gliding on this surface (see Fig. 3.1, right).

Poincaré Cuts. We fix first the energy H_0 and put $q_{10} = 0$. Then for any point $P_0 = (q_{20}, p_{20})$, we obtain p_{10} from (3.1) as $p_{10} = \sqrt{2H_0 - 2U_0 - p_{20}^2}$, where we choose the positive root. We then follow the solution until it hits again the surface $q_1 = 0$ in the positive direction $p_1 > 0$ and obtain a point $P_1 = (q_{21}, p_{21})$; in the same way we compute $P_2 = (q_{22}, p_{22})$, etc. For the same initial values as in Fig. 3.1 and with $H_0 = \frac{1}{12}$, the solution for $0 \le t \le 300\,000$ gives $46\,865$ Poincaré cuts which are all displayed in Fig. 3.2 (left). They seem to lie exactly on a curve, as do the orbits for 5 other choices of initial values. This picture thus shows "convincing

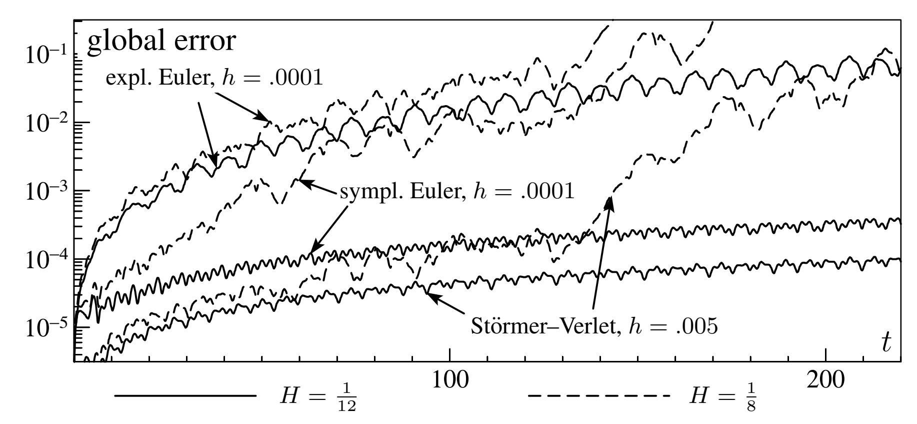


Fig. 3.4. Global error of numerical methods for nearly quasiperiodic and for chaotic solutions; same initial data as in Fig. 3.2

evidence" for the existence of a second invariant, for which Gustavson (1966) has derived a formal expansion, whose first terms represent perfectly these curves.

"But here comes the surprise" (Hénon–Heiles, p. 76): Fig. 3.2 shows to the right the same picture in the (q_2, p_2) plane for a somewhat higher Energy $H = \frac{1}{8}$. The motion turns completely to chaos and all hope for a second invariant disappears. Actually, Gustavson's series does not converge.

Numerical Experiments. We now apply numerical methods, the *explicit* Euler method to the low energy initial values $H = \frac{1}{12}$ (Fig. 3.3, left), and the *implicit* Euler method to the high energy initial values (Fig. 3.3, right), both methods with a very small step size $h = 10^{-5}$. As we already expect from our previous experiences, the explicit Euler method tends to *increase* the energy and turns order into chaos, while the implicit Euler method tends to *decrease* it and turns chaos into order. The Störmer–Verlet method (not shown) behaves as the exact solution even for step sizes as large as $h = 10^{-1}$.

In our next experiment we study the *global error* (see Fig. 3.4), once for the case of the nearly quasiperiodic orbit ($H = \frac{1}{12}$) and once for the chaotic one ($H = \frac{1}{8}$), both for the explicit Euler, the symplectic Euler, and the Störmer–Verlet scheme. It may come as a surprise, that only in the first case we have the same behaviour (linear or quadratic growth) as in Fig. 2.3 for the Kepler problem. In the second case ($H = \frac{1}{8}$) the global error grows exponentially for all methods, and the explicit Euler method is worst.

Study of a Mapping. The passage from a point P_i to the next one P_{i+1} (as explained for the left picture of Fig. 3.2) can be considered as a *mapping* $\Phi : P_i \mapsto P_{i+1}$ and the sequence of points $P_0, P_1, P_2, \ldots$ are just the iterates of this mapping. This mapping is represented for the two energy levels $H = \frac{1}{12}$ and $H = \frac{1}{8}$ in Fig. 3.5 and its study allows to better understand the behaviour of the orbits. We see no significant difference between the two cases, simply for larger H the deformations are more violent and correspond to larger eigenvalues of the Jacobian of Φ. In

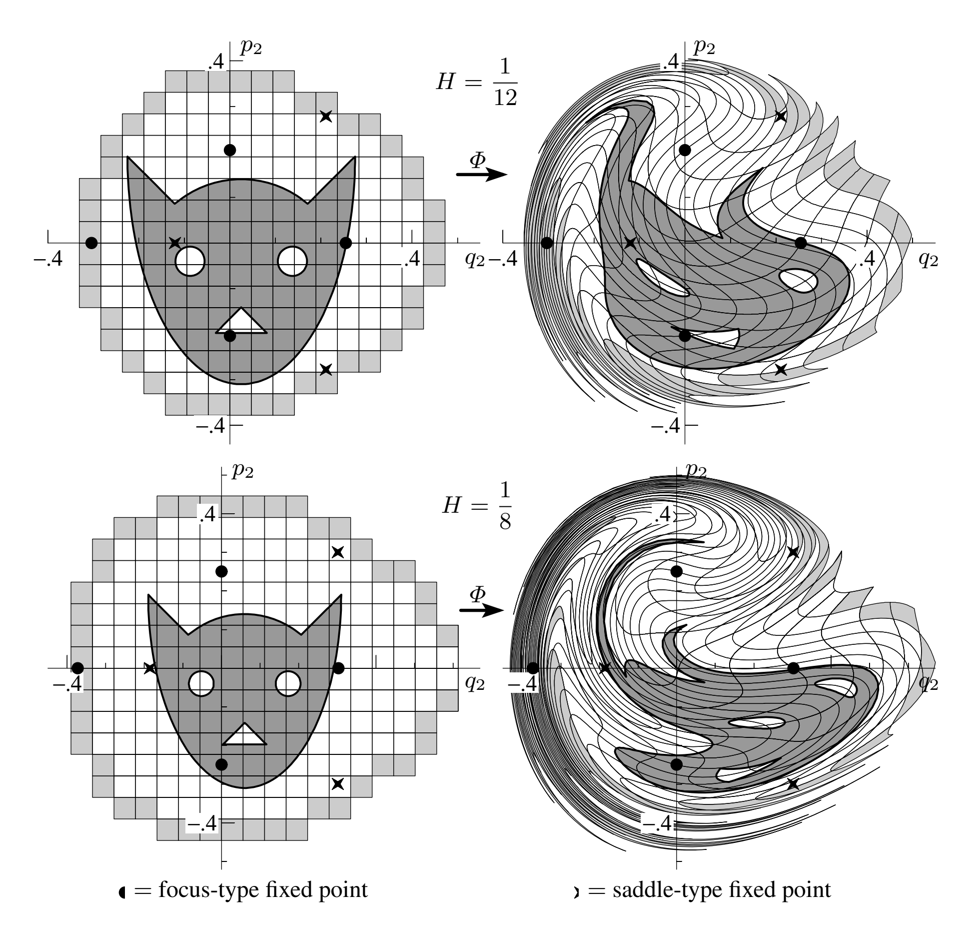


Fig. 3.5. The Poincaré map $\Phi : P_0 \to P_1$ for the Hénon–Heiles Model

both cases we have seven fixed points, which correspond to periodic solutions of the system (3.3). Four of them are stable and lie inside the white islands of Fig. 3.2.

I.4 Molecular Dynamics

> We do not need exact classical trajectories to do this, but must lay great emphasis on energy conservation as being of primary importance for this reason. (M.P. Allen & D.J. Tildesley 1987)

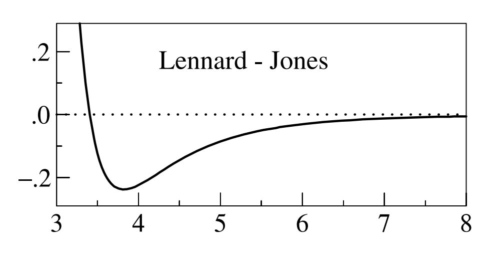
Molecular dynamics requires the solution of Hamiltonian systems (1.10), where the total energy is given by

$$H(p,q) = \frac{1}{2}\sum_{i=1}^{N} \frac{1}{m_i}\, p_i^T p_i + \sum_{i=2}^{N}\sum_{j=1}^{i-1} V_{ij}\Big(\|q_i - q_j\|\Big), \tag{4.1}$$

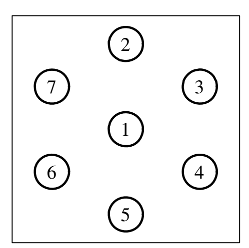
and $V_{ij}(r)$ are given potential functions. Here, q_i and p_i denote the positions and momenta of atoms and m_i is the atomic mass of the ith atom. We remark that the outer solar system (2.12) is such an N-body system with $V_{ij}(r) = -Gm_im_j/r$. In molecular dynamics the Lennard–Jones potential

$$V_{ij}(r) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right) \tag{4.2}$$

is very popular (ε_{ij} and σ_{ij} are suitable constants depending on the atoms). This potential has an absolute minimum at distance $r = \sigma_{ij}\sqrt[6]{2}$. The force due to this potential strongly repels the atoms when they are closer than this value, and they attract each other when they are farther away.



Numerical Experiments with a Frozen Argon Crystal. As in Biesiadecki & Skeel (1993) we consider the interaction of seven argon atoms in a plane, where six of them are arranged symmetrically around a centre atom. As a mathematical model we take the Hamiltonian (4.1) with $N = 7$, $m_i = m = 66.34 \cdot 10^{-27}$ [kg],



$$\varepsilon_{ij} = \varepsilon = 119.8\, k_B \text{ [J]}, \qquad \sigma_{ij} = \sigma = 0.341 \text{ [nm]},$$

where $k_B = 1.380658 \cdot 10^{-23}$ [J/K] is Boltzmann's constant (see Allen & Tildesley (1987), page 21). As units for our calculations we take masses in [kg], distances in nanometers (1 [nm] $= 10^{-9}$ [m]), and times in nanoseconds (1 [nsec] $= 10^{-9}$ [sec]). Initial positions (in [nm]) and initial velocities (in [nm/nsec]) are given in Table 4.1. They are chosen such that neighbouring atoms have a distance that is close to the one with lowest potential energy, and such that the total momentum is zero and therefore the centre of gravity does not move. The energy at the initial position is $H(p_0, q_0) \approx -1260.2\, k_B$ [J].

For computations in molecular dynamics one is usually not interested in the trajectories of the atoms, but one aims at macroscopic quantities such as temperature, pressure, internal energy, etc. Here we consider the total energy, given by the Hamiltonian, and the temperature which can be calculated from the formula (see Allen &

Table 4.1. Initial values for the simulation of a frozen argon crystal

atom	1	2	3	4	5	6	7
position	0.00 0.00	0.02 0.39	0.34 0.17	0.36 −0.21	−0.02 −0.40	−0.35 −0.16	−0.31 0.21
velocity	−30 −20	50 −90	−70 −60	90 40	80 90	−40 100	−80 −60

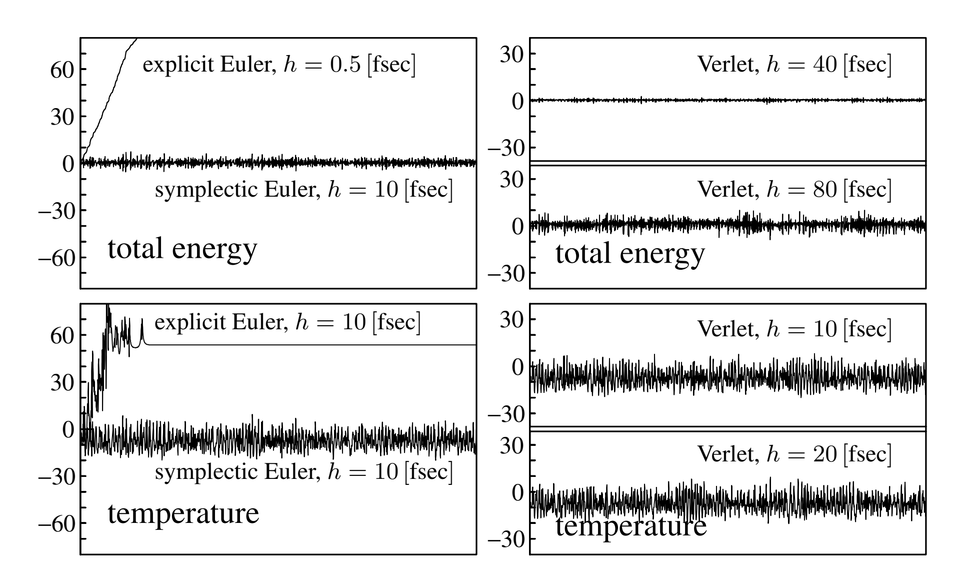


Fig. 4.1. Computed total energy and temperature of the argon crystal

Tildesley (1987), page 46)

$$T = \frac{1}{2Nk_B} \sum_{i=1}^{N} m_i \|\dot{q}_i\|^2. \tag{4.3}$$

We apply the explicit and symplectic Euler methods and also the Verlet method to this problem. Observe that for a Hamiltonian such as (4.1) all three methods are explicit, and all of them need only one force evaluation per integration step. In Fig. 4.1 we present the numerical results of our experiments. The integrations are done over an interval of length 0.2 [nsec]. The step sizes are indicated in femtoseconds (1 [fsec] $= 10^{-6}$ [nsec]).

The two upper pictures show the values $\big(H(p_n, q_n) - H(p_0, q_0)\big)/k_B$ as a function of time $t_n = nh$. For the exact solution, this value is precisely zero for all times. Similar to earlier experiments we see that the symplectic Euler method is qualitatively correct, whereas the numerical solution of the explicit Euler method, although computed with a much smaller step size, is completely useless (see the citation at the beginning of this section). The Verlet method is qualitatively correct and gives much more accurate results than the symplectic Euler method (we shall see later that the Verlet method is of order 2). The two computations with the Verlet method show that the energy error decreases by a factor of 4 if the step size is reduced by a factor of 2 (second order convergence).

The two lower pictures of Fig. 4.1 show the numerical values of the temperature difference $T - T_0$ with T given by (4.3) and $T_0 \approx 22.72$ [K] (initial temperature). In contrast to the total energy, this is not an exact invariant, but for our problem it fluctuates around a constant value. The explicit Euler method gives wrong results,

but the symplectic Euler and the Verlet methods show the desired behaviour. This time a reduction of the step size does not reduce the amplitude of the oscillations, which indicates that the fluctuation of the exact temperature is of the same size.

I.5 Highly Oscillatory Problems

In this section we discuss a system with almost-harmonic high-frequency oscillations. We show numerical phenomena of methods applied with step sizes that are not small compared to the period of the fastest oscillations.

I.5.1 A Fermi–Pasta–Ulam Problem

> . . . dealing with the behavior of certain nonlinear physical systems where the non-linearity is introduced as a perturbation to a primarily linear problem. The behavior of the systems is to be studied for times which are long compared to the characteristic periods of the corresponding linear problems. (E. Fermi, J. Pasta, S. Ulam 1955)

> In the early 1950s MANIAC-I had just been completed and sat poised for an attack on significant problems. ... Fermi suggested that it would be highly instructive to integrate the equations of motion numerically for a judiciously chosen, one-dimensional, harmonic chain of mass points weakly perturbed by nonlinear forces. (J. Ford 1992)

The problem of Fermi, Pasta & Ulam (1955) is a simple model for simulations in statistical mechanics which revealed highly unexpected dynamical behaviour. We consider a modification consisting of a chain of $2m$ mass points, connected with alternating soft nonlinear and stiff linear springs, and fixed at the end points (see Galgani, Giorgilli, Martinoli & Vanzini (1992) and Fig. 5.1). The variables $q_1, \ldots, q_{2m}$

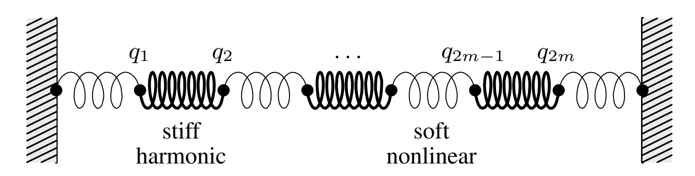


Fig. 5.1. Chain with alternating soft nonlinear and stiff linear springs

($q_0 = q_{2m+1} = 0$) stand for the displacements of the mass points, and $p_i = \dot{q}_i$ for their velocities. The motion is described by a Hamiltonian system with total energy

$$H(p,q) = \frac{1}{2}\sum_{i=1}^{m}\big(p_{2i-1}^2 + p_{2i}^2\big) + \frac{\omega^2}{4}\sum_{i=1}^{m}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{m}(q_{2i+1} - q_{2i})^4,$$

where ω is assumed to be large. It is quite natural to introduce the new variables

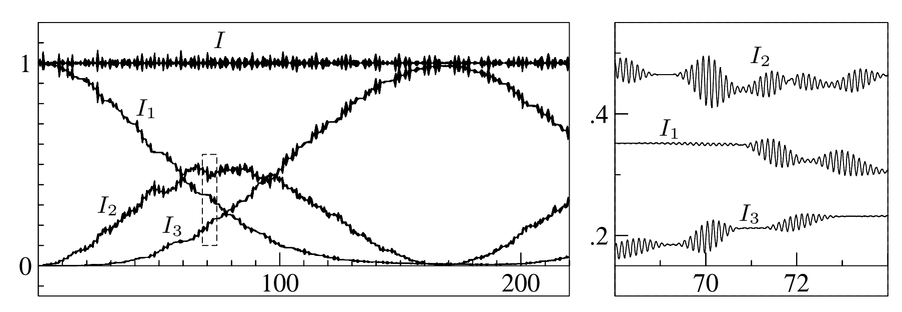


Fig. 5.2. Exchange of energy in the exact solution of the Fermi–Pasta–Ulam model. The picture to the right is an enlargement of the narrow rectangle in the left-hand picture

$$\begin{aligned} x_{0,i} &= \big(q_{2i}+q_{2i-1}\big)/\sqrt{2}, \qquad & x_{1,i} &= \big(q_{2i}-q_{2i-1}\big)/\sqrt{2}, \\ y_{0,i} &= \big(p_{2i}+p_{2i-1}\big)/\sqrt{2}, \qquad & y_{1,i} &= \big(p_{2i}-p_{2i-1}\big)/\sqrt{2}, \end{aligned} \tag{5.1}$$

where $x_{0,i}$ ($i = 1,\ldots,m$) represents a scaled displacement of the ith stiff spring, $x_{1,i}$ a scaled expansion (or compression) of the ith stiff spring, and $y_{0,i}, y_{1,i}$ their velocities (or momenta). With this change of coordinates, the motion in the new variables is again described by a Hamiltonian system, with

$$\begin{aligned} H(y,x) &= \frac{1}{2}\sum_{i=1}^{m}\Big(y_{0,i}^2+y_{1,i}^2\Big)+\frac{\omega^2}{2}\sum_{i=1}^{m}x_{1,i}^2+\frac{1}{4}\Big((x_{0,1}-x_{1,1})^4+ \\ &\quad +\sum_{i=1}^{m-1}\big(x_{0,i+1}-x_{1,i+1}-x_{0,i}-x_{1,i}\big)^4+(x_{0,m}+x_{1,m})^4\Big). \end{aligned} \tag{5.2}$$

Besides the fact that the equations of motion are Hamiltonian, so that the total energy is exactly conserved, they have a further interesting feature. Let

$$I_j(x_{1,j},y_{1,j}) = \frac{1}{2}\big(y_{1,j}^2+\omega^2 x_{1,j}^2\big) \tag{5.3}$$

denote the energy of the jth stiff spring. It turns out that there is an exchange of energy between the stiff springs, but the total oscillatory energy $I = I_1 + \ldots + I_m$ remains close to a constant value, in fact, $I\big((x(t),y(t)\big) = I\big((x(0),y(0)\big) + \mathcal{O}(\omega^{-1})$. For an illustration of this property, we choose $m = 3$ (as in Fig. 5.1), $\omega = 50$,

$$x_{0,1}(0)=1, \quad y_{0,1}(0)=1, \quad x_{1,1}(0)=\omega^{-1}, \quad y_{1,1}(0)=1,$$

and zero for the remaining initial values. Fig. 5.2 displays the energies I_1, I_2, I_3 of the stiff springs together with the total oscillatory energy $I = I_1 + I_2 + I_3$ as a function of time. The solution has been computed very carefully with high accuracy, so that the displayed oscillations can be considered as exact.

I.5.2 Application of Classical Integrators

Which of the methods of the foregoing sections produce qualitatively correct approximations when the product of the step size h with the high frequency ω is relatively large?

Linear Stability Analysis. To get an idea of the maximum admissible step size, we neglect the quartic term in the Hamiltonian (5.2), so that the differential equation splits into the two-dimensional problems $\dot{y}_{0,i} = 0$, $\dot{x}_{0,i} = y_{0,i}$ and

$$\dot{y}_{1,i} = -\omega^2 x_{1,i}, \quad \dot{x}_{1,i} = y_{1,i}. \tag{5.4}$$

Omitting the subscripts, the solution of (5.4) is

$$\begin{pmatrix} y(t) \\ \omega x(t) \end{pmatrix} = \begin{pmatrix} \cos \omega t & -\sin \omega t \\ \sin \omega t & \cos \omega t \end{pmatrix} \begin{pmatrix} y(0) \\ \omega x(0) \end{pmatrix}.$$

The numerical solution of a one-step method applied to (5.4) yields

$$\begin{pmatrix} y_{n+1} \\ \omega x_{n+1} \end{pmatrix} = M(h\omega) \begin{pmatrix} y_n \\ \omega x_n \end{pmatrix}, \tag{5.5}$$

and the eigenvalues λ_i of $M(h\omega)$ determine the long-time behaviour of the numerical solution. Stability (i.e., boundedness of the solution of (5.5)) requires the eigenvalues to be less than or equal to one in modulus. For the explicit Euler method we have $\lambda_{1,2} = 1 \pm ih\omega$, so that the energy $I_n = (y_n^2 + \omega^2 x_n^2)/2$ increases as $(1 + h^2\omega^2)^{n/2}$. For the implicit Euler method we have $\lambda_{1,2} = (1 \pm ih\omega)^{-1}$, and the energy decreases as $(1 + h^2\omega^2)^{-n/2}$. For the implicit midpoint rule, the matrix $M(h\omega)$ is orthogonal and therefore I_n is exactly preserved for all h and for all times. Finally, for the symplectic Euler method and for the Störmer–Verlet scheme we have

$$M(h\omega) = \begin{pmatrix} 1 & -h\omega \\ h\omega & 1 - h^2\omega^2 \end{pmatrix}, \quad M(h\omega) = \begin{pmatrix} 1 - \frac{h^2\omega^2}{2} & -\frac{h\omega}{2}\Big(1 - \frac{h^2\omega^2}{4}\Big) \\ \frac{h\omega}{2} & 1 - \frac{h^2\omega^2}{2} \end{pmatrix},$$

respectively. For both matrices, the characteristic polynomial is $\lambda^2 - (2 - h^2\omega^2)\lambda + 1$, so that the eigenvalues are of modulus one if and only if $|h\omega| \leq 2$.

Numerical Experiments. We apply several methods to the Fermi–Pasta–Ulam (FPU) problem, with $\omega = 50$ and initial data as given in Sect. I.5.1. The explicit and implicit Euler methods give completely wrong solutions even for very small step sizes. Fig. 5.3 presents the numerical results for H, I, I_1, I_2, I_3 obtained with the implicit midpoint rule, the symplectic Euler, and the Störmer–Verlet scheme. For the small step size $h = 0.001$ all methods give satisfactory results, although the energy exchange is not reproduced accurately over long times. The Hamiltonian H and the total oscillatory energy I are well conserved over much longer time intervals. The larger step size $h = 0.03$ has been chosen such that $h\omega = 1.5$ is close

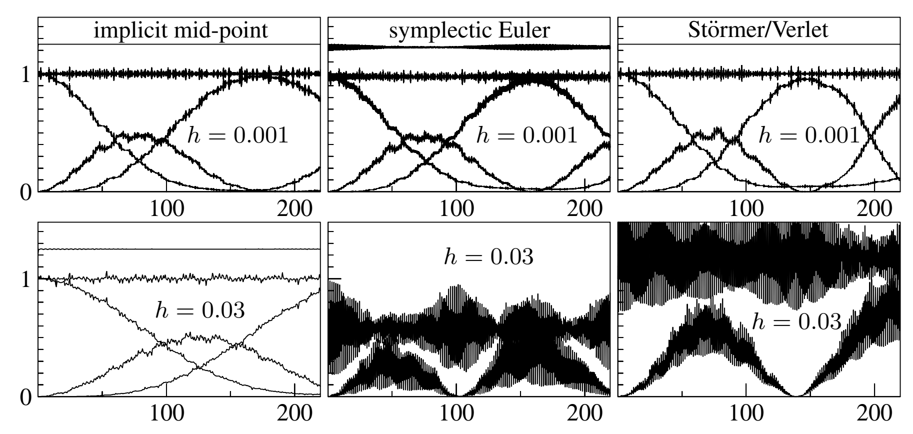


Fig. 5.3. Numerical solution for the FPU problem (5.2) with data as in Sect. I.5.1, obtained with the implicit midpoint rule (left), symplectic Euler (middle), and Störmer–Verlet scheme (right); the upper pictures use $h = 0.001$, the lower pictures $h = 0.03$; the first four pictures show the Hamiltonian $H - 0.8$ and the oscillatory energies I_1, I_2, I_3, I; the last two pictures only show I_2 and I

to the stability limit of the symplectic Euler and the Störmer–Verlet methods. The values of H and I are still bounded over very long time intervals, but the oscillations do not represent the true behaviour. Moreover, the average value of I is no longer close to 1, as it is for the exact solution. These phenomena call for an explanation, and for numerical methods with an improved behaviour (see Chap. XIII).

I.6 Exercises

1. Show that the Lotka–Volterra problem (1.1) in logarithmic scale, i.e., by putting $p = \log u$ and $q = \log v$, becomes a Hamiltonian system with the function (1.4) as Hamiltonian (see Fig. 6.1).

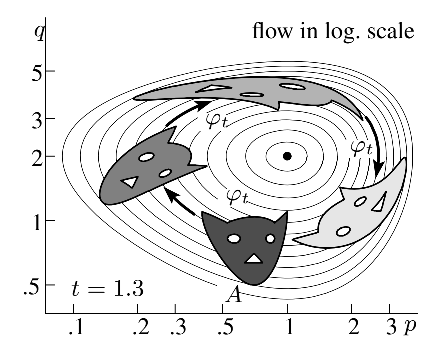


Fig. 6.1. Area preservation in logarithmic scale of the Lotka–Volterra flow

2. Apply the symplectic Euler method (or the implicit midpoint rule) to problems such as

$$\begin{pmatrix} \dot u \\ \dot v \end{pmatrix} = \begin{pmatrix} (v-2)/v \\ (1-u)/u \end{pmatrix}, \qquad \begin{pmatrix} \dot u \\ \dot v \end{pmatrix} = \begin{pmatrix} u^2 v(v-2) \\ v^2 u(1-u) \end{pmatrix}$$

with various initial conditions. Both problems have the same first integral (1.4) as the Lotka–Volterra problem and therefore their solutions are also periodic. Do the numerical solutions also show this behaviour?

3. A general two-body problem (sun and planet) is given by the Hamiltonian

$$H(p, p_S, q, q_S) = \frac{1}{2M}\, p_S^T p_S + \frac{1}{2m}\, p^T p - \frac{GmM}{\|q - q_S\|},$$

where $q_S, q \in \mathbb{R}^3$ are the positions of the sun (mass M) and the planet (mass m), $p_S, p \in \mathbb{R}^3$ are their momenta, and G is the gravitational constant.

a) Prove: in heliocentric coordinates $Q := q - q_S$, the equations of motion are

$$\ddot Q = -G(M+m)\,\frac{Q}{\|Q\|^3}.$$

b) Prove that $\frac{d}{dt}\big(Q(t) \times \dot Q(t)\big) = 0$, so that $Q(t)$ stays for all times t in the plane $E = \{q\,;\, d^T q = 0\}$, where $d = Q(0) \times \dot Q(0)$.

Conclusion. The coordinates corresponding to a basis in E satisfy the two-dimensional equations (2.2).

4. In polar coordinates, the two-body problem (2.2) becomes

$$\ddot r = -V'(r) \qquad \text{with} \qquad V(r) = \frac{L_0^2}{2r^2} - \frac{1}{r}$$

which is independent of φ. The angle $\varphi(t)$ can be obtained by simple integration from $\dot\varphi(t) = L_0/r^2(t)$.

5. Compute the period of the solution of the Kepler problem (2.2) and deduce from the result Kepler's "third law".

Hint. Comparing Kepler's second law (2.6) with the area of the ellipse gives $\frac{1}{2}L_0 T = ab\pi$. Then apply (2.7). The result is $T = 2\pi(2|H_0|)^{-3/2} = 2\pi a^{3/2}$.

6. Deduce Kepler's first law from (2.2) by the elegant method of Laplace (1799).

Hint. Multiplying (2.2) with (2.5) gives

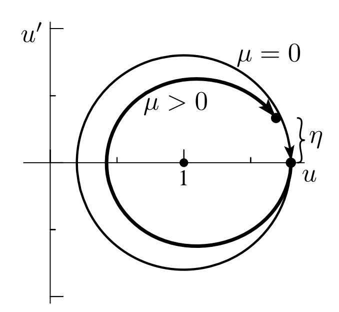
$$L_0 \ddot q_1 = \frac{d}{dt}\Big(\frac{q_2}{r}\Big), \qquad L_0 \ddot q_2 = \frac{d}{dt}\Big(-\frac{q_1}{r}\Big),$$

and after integration $L_0 \dot q_1 = \frac{q_2}{r} + B$, $L_0 \dot q_2 = -\frac{q_1}{r} + A$, where A and B are integration constants. Then eliminate $\dot q_1$ and $\dot q_2$ by multiplying these equations by q_2 and $-q_1$ respectively and by subtracting them. The result is a quadratic equation in q_1 and q_2.

7. Whatever the initial values for the Kepler problem are, $1 + 2H_0 L_0^2 \ge 0$ holds. Hence, the value e is well defined by (2.9).

Hint. L_0 is the area of the parallelogram spanned by the vectors $q(0)$ and $\dot q(0)$.

8. *Implementation of the Störmer–Verlet scheme.* Explain why the use of the one-step formulation (1.17) is numerically more stable than that of the two-term recursion (1.15).
9. *Runge–Lenz–Pauli vector.* Prove that the function
$$A(p,q) = \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ q_1p_2 - q_2p_1 \end{pmatrix} - \frac{1}{\sqrt{q_1^2+q_2^2}} \begin{pmatrix} q_1 \\ q_2 \\ 0 \end{pmatrix}$$
is a first integral of the Kepler problem, i.e., $A\big(p(t),q(t)\big) = Const$ along solutions of the problem. However, it is not a first integral of the perturbed Kepler problem of Exercise 12.
10. Add a column to Table 2.1 which shows the long-time behaviour of the error in the Runge–Lenz–Pauli vector (see Exercise 9) for the various numerical integrators.
11. For the Kepler problem, eliminate (p_1,p_2) from the relations $H(p,q) = Const$, $L(p,q) = Const$ and $A(p,q) = Const$. This gives a quadratic relation for (q_1,q_2) and proves that the solution lies on an ellipse, a parabola, or on a hyperbola.
12. Study numerically the solution of the perturbed Kepler problem with Hamiltonian
$$H(p_1,p_2,q_1,q_2) = \frac{1}{2}\left(p_1^2+p_2^2\right) - \frac{1}{\sqrt{q_1^2+q_2^2}} - \frac{\mu}{3\sqrt{(q_1^2+q_2^2)^3}},$$
where μ is a positive or negative small number. Among others, this problem describes the motion of a planet in the Schwarzschild potential for Einstein's general relativity theory[7]. You will observe a precession of the perihelion, which, applied to the orbit of Mercury, represented the historically first verification of Einstein's theory (see e.g., Birkhoff 1923, p. 261-264).



The precession can also be expressed analytically: the equation for $u = 1/r$ as a function of φ, corresponding to (2.8), here becomes
$$u'' + u = \frac{1}{d} + \mu u^2, \tag{6.1}$$
where $d = L_0^2$. Now compute the derivative of this solution with respect to μ, at $\mu = 0$ and $u = (1 + e\cos(\varphi - \varphi^*))/d$ after one period $t = 2\pi$. This leads to $\eta = \mu(e/d^2)\cdot 2\pi\sin\varphi$ (see the small picture). Then, for small μ, the precession after one period is
$$\Delta\varphi = \frac{2\pi\mu}{d}. \tag{6.2}$$

[7] We are grateful to Prof. Ruth Durrer for helpful hints about this subject.

Chapter II.
Numerical Integrators

After having seen in Chap. I some simple numerical methods and a variety of numerical phenomena that they exhibited, we now present more elaborate classes of numerical methods. We start with Runge–Kutta and collocation methods, and we introduce discontinuous collocation methods, which cover essentially all high-order implicit Runge–Kutta methods of interest. We then treat partitioned Runge–Kutta methods and Nyström methods, which can be applied to partitioned problems such as Hamiltonian systems. Finally we present composition and splitting methods.

II.1 Runge–Kutta and Collocation Methods



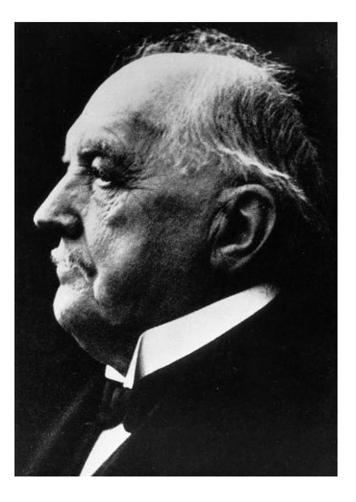


Fig. 1.1. Carl David Tolmé Runge (left picture), born: 30 August 1856 in Bremen (Germany), died: 3 January 1927 in Göttingen (Germany).
Wilhelm Martin Kutta (right picture), born: 3 November 1867 in Pitschen, Upper Silesia (now Byczyna, Poland), died: 25 December 1944 in Fürstenfeldbruck (Germany)

Runge–Kutta methods form an important class of methods for the integration of differential equations. A special subclass, the collocation methods, allows for a particularly elegant access to order, symplecticity and continuous output.

II.1.1 Runge–Kutta Methods

In this section, we treat non-autonomous systems of first-order ordinary differential equations

$$\dot{y} = f(t,y), \qquad y(t_0) = y_0. \tag{1.1}$$

The integration of this equation gives $y(t_1) = y_0 + \int_{t_0}^{t_1} f(t,y(t))\,dt$, and replacing the integral by the trapezoidal rule, we obtain

$$y_1 = y_0 + \frac{h}{2}\big(f(t_0,y_0) + f(t_1,y_1)\big). \tag{1.2}$$

This is the *implicit trapezoidal rule*, which, in addition to its historical importance for computations in partial differential equations (Crank–Nicolson) and in A-stability theory (Dahlquist), played a crucial role even earlier in the discovery of Runge–Kutta methods. It was the starting point of Runge (1895), who "predicted" the unknown y_1-value to the right by an Euler step, and obtained the first of the following formulas (the second being the analogous formula for the midpoint rule)

$$\begin{aligned} k_1 &= f(t_0,y_0) & k_1 &= f(t_0,y_0) \\ k_2 &= f(t_0+h, y_0+hk_1) & k_2 &= f(t_0+\tfrac{h}{2}, y_0+\tfrac{h}{2}k_1) \\ y_1 &= y_0 + \tfrac{h}{2}\big(k_1+k_2\big) & y_1 &= y_0 + hk_2. \end{aligned} \tag{1.3}$$

These methods have a nice geometric interpretation (which is illustrated in the first two pictures of Fig. 1.2 for a famous problem, the Riccati equation): they consist of polygonal lines, which assume the slopes prescribed by the differential equation evaluated at previous points.

Idea of Heun (1900) and Kutta (1901): compute *several* polygonal lines, each starting at y_0 and assuming the various slopes k_j on portions of the integration interval, which are proportional to some given constants a_{ij}; at the final point of each polygon evaluate a new slope k_i. The last of these polygons, with constants b_i, determines the numerical solution y_1 (see the third picture of Fig. 1.2). This idea leads to the class of *explicit* Runge–Kutta methods, i.e., formula (1.4) below with $a_{ij} = 0$ for $i \le j$.

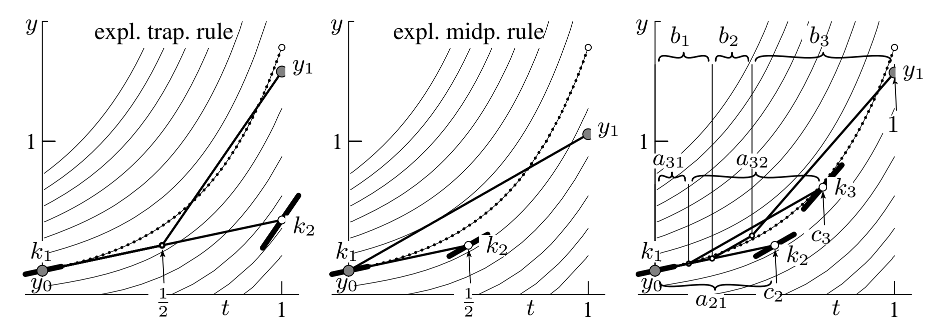


Fig. 1.2. Runge–Kutta methods for $\dot{y} = t^2 + y^2$, $y_0 = 0.46$, $h = 1$; dotted: exact solution

Much more important for our purpose are *implicit* Runge–Kutta methods, introduced mainly in the work of Butcher (1963).

Definition 1.1. Let b_i, a_{ij} $(i, j = 1, \ldots, s)$ be real numbers and let $c_i = \sum_{j=1}^{s} a_{ij}$. An *s-stage Runge–Kutta method* is given by

$$\begin{aligned} k_i &= f\Big(t_0 + c_i h, y_0 + h \sum_{j=1}^{s} a_{ij} k_j\Big), \quad i = 1, \ldots, s \\ y_1 &= y_0 + h \sum_{i=1}^{s} b_i k_i. \end{aligned} \tag{1.4}$$

Here we allow a full matrix (a_{ij}) of non-zero coefficients. In this case, the slopes k_i can no longer be computed explicitly, and even do not necessarily exist. For example, for the problem set-up of Fig. 1.2 the implicit trapezoidal rule has no solution. However, the implicit function theorem assures that, for sufficiently small h, the nonlinear system (1.4) for the values $k_1, \ldots, k_s$ has a locally unique solution close to $k_i \approx f(t_0, y_0)$.

Since Butcher's work, the coefficients are usually displayed as follows:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \,. \tag{1.5}$$

Definition 1.2. A Runge–Kutta method (or a general one-step method) has *order* p, if for all sufficiently regular problems (1.1) the *local error* $y_1 - y(t_0 + h)$ satisfies

$$y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1}) \quad \text{as } h \to 0.$$

To check the order of a Runge Kutta method, one has to compute the Taylor series expansions of $y(t_0 + h)$ and y_1 around to $h = 0$. This leads to the following algebraic conditions for the coefficients for orders 1, 2, and 3:

$$\begin{array}{rrl} & \sum_i b_i = 1 & \text{for order 1;} \\ \text{in addition} & \sum_i b_i c_i = 1/2 & \text{for order 2;} \\ \text{in addition} & \sum_i b_i c_i^2 = 1/3 & \\ \text{and} & \sum_{i,j} b_i a_{ij} c_j = 1/6 & \text{for order 3.} \end{array} \tag{1.6}$$

For higher orders, however, this problem represented a great challenge in the first half of the 20th century. We shall present an elegant theory in Sect. III.1 which allows order conditions to be derived.

Among the methods seen up to now, the explicit and implicit Euler methods

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \qquad\qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{1.7}$$

are of order 1, the implicit trapezoidal and midpoint rules as well as both methods of Runge

$$
\begin{array}{c|cc}
0 & & \\
1 & 1/2 & 1/2 \\
\hline
 & 1/2 & 1/2
\end{array}
\qquad
\begin{array}{c|c}
1/2 & 1/2 \\
\hline
 & 1
\end{array}
\qquad
\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
 & 1/2 & 1/2
\end{array}
\qquad
\begin{array}{c|cc}
0 & & \\
1/2 & 1/2 & \\
\hline
 & 0 & 1
\end{array}
$$

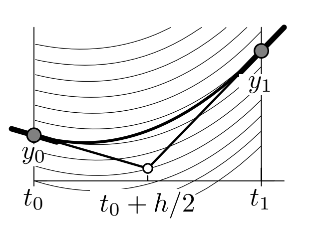
are of order 2. The most successful methods during more than half a century were the 4th order methods of Kutta:

$$
\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
1/2 & 0 & 1/2 & & \\
1 & 0 & 0 & 1 & \\
\hline
 & 1/6 & 2/6 & 2/6 & 1/6
\end{array}
\qquad\qquad
\begin{array}{c|cccc}
0 & & & & \\
1/3 & 1/3 & & & \\
2/3 & -1/3 & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
 & 1/8 & 3/8 & 3/8 & 1/8
\end{array}
\tag{1.8}
$$

II.1.2 Collocation Methods

> The high speed computing machines make it possible to enjoy the advantages of intricate methods. (P.C. Hammer & J.W. Hollingsworth 1955)

Collocation methods for ordinary differential equations have their origin, once again, in the implicit trapezoidal rule (1.2): Hammer & Hollingsworth (1955) discovered that this method can be interpreted as being generated by a *quadratic function* "which agrees in direction with that indicated by the differential equation at two points" t_0 and t_1 (see the picture to the right). This idea allows one to "see much-used methods in a new light" and allows various generalizations (Guillou & Soulé (1969), Wright (1970)). An interesting feature of collocation methods is that we not only get a discrete set of approximations, but also a *continuous approximation* to the solution.



Definition 1.3. Let $c_1, \ldots, c_s$ be distinct real numbers (usually $0 \le c_i \le 1$). The *collocation polynomial* $u(t)$ is a polynomial of degree s satisfying

$$
\begin{aligned}
u(t_0) &= y_0 \\
\dot u(t_0 + c_i h) &= f\big(t_0 + c_i h, u(t_0 + c_i h)\big), \quad i = 1, \ldots, s,
\end{aligned}
\tag{1.9}
$$

and the numerical solution of the *collocation method* is defined by $y_1 = u(t_0 + h)$.

For $s = 1$, the polynomial has to be of the form $u(t) = y_0 + (t - t_0)k$ with

$$k = f(t_0 + c_1 h, y_0 + h c_1 k).$$

We see that the explicit and implicit Euler methods and the midpoint rule are collocation methods with $c_1 = 0$, $c_1 = 1$ and $c_1 = 1/2$, respectively.

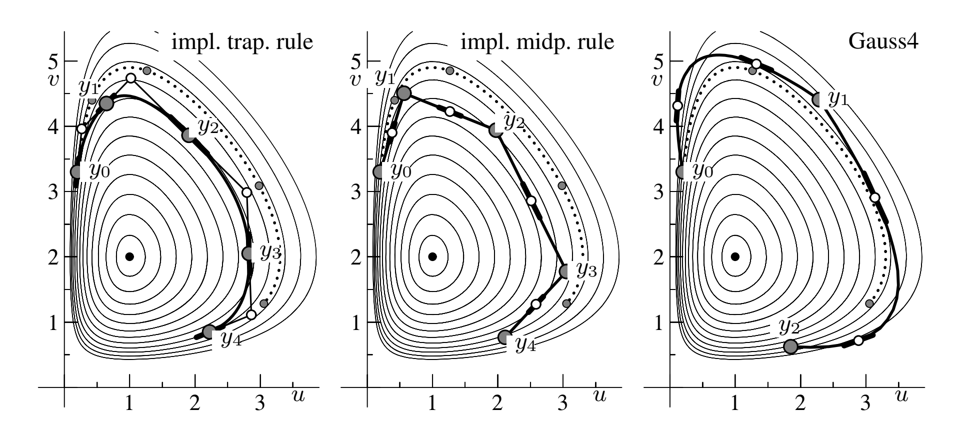
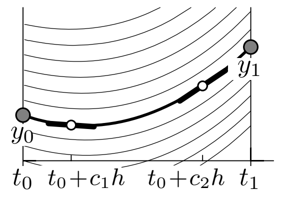


Fig. 1.3. Collocation solutions for the Lotka–Volterra problem (I.1.1); $u_0 = 0.2$, $v_0 = 3.3$; methods of order 2: four steps with $h = 0.4$; method of order 4: two steps with $h = 0.8$; dotted: exact solution

For $s = 2$ and $c_1 = 0, c_2 = 1$ we find, of course, the implicit trapezoidal rule. The choice of Hammer & Hollingsworth for the collocation points is $c_{1,2} = 1/2 \pm \sqrt{3}/6$, the *Gaussian quadrature nodes* (see the picture to the right). We will see that the corresponding method is of order 4.



In Fig. 1.3 we illustrate the collocation idea with these methods for the Lotka–Volterra problem (I.1.1). One can observe that, in spite of the extremely large step sizes, the methods are quite satisfactory.

Theorem 1.4 (Guillou & Soulé 1969, Wright 1970). *The collocation method of Definition 1.3 is equivalent to the s-stage Runge–Kutta method (1.4) with coefficients*

$$a_{ij} = \int_0^{c_i} \ell_j(\tau)\, d\tau, \qquad b_i = \int_0^1 \ell_i(\tau)\, d\tau, \tag{1.10}$$

where $\ell_i(\tau)$ is the Lagrange polynomial $\ell_i(\tau) = \prod_{l \neq i} (\tau - c_l)/(c_i - c_l)$.

Proof. Let $u(t)$ be the collocation polynomial and define

$$k_i := \dot{u}(t_0 + c_i h).$$

By the Lagrange interpolation formula we have $\dot{u}(t_0 + \tau h) = \sum_{j=1}^{s} k_j \cdot \ell_j(\tau)$, and by integration we get

$$u(t_0 + c_i h) = y_0 + h \sum_{j=1}^{s} k_j \int_0^{c_i} \ell_j(\tau)\, d\tau.$$

Inserted into (1.9) this gives the first formula of the Runge–Kutta equation (1.4). Integration from 0 to 1 yields the second one. □

The above proof can also be read in reverse order. This shows that a Runge–Kutta method with coefficients given by (1.10) can be interpreted as a collocation method. Since $\tau^{k-1} = \sum_{j=1}^{s} c_j^{k-1} \ell_j(\tau)$ for $k = 1, \ldots, s$, the relations (1.10) are equivalent to the linear systems

$$\begin{aligned} C(q): \qquad & \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \quad k = 1, \ldots, q, \ \text{all } i \\ B(p): \qquad & \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p, \end{aligned} \tag{1.11}$$

with $q = s$ and $p = s$. What is the order of a Runge–Kutta method whose coefficients b_i, a_{ij} are determined in this way?

Compared to the enormous difficulties that the first explorers had in constructing Runge–Kutta methods of orders 5 and 6, and also compared to the difficult algebraic proofs of the first papers of Butcher, the following general theorem and its proof, discovered in this form by Guillou & Soulé (1969), are surprisingly simple.

Theorem 1.5 (Superconvergence). *If the condition $B(p)$ holds for some $p \geq s$, then the collocation method (Definition 1.3) has order p. This means that the collocation method has the same order as the underlying quadrature formula.*

Proof. We consider the collocation polynomial $u(t)$ as the solution of a perturbed differential equation

$$\dot{u} = f(t, u) + \delta(t) \tag{1.12}$$

with defect $\delta(t) := \dot{u}(t) - f\big(t, u(t)\big)$. Subtracting (1.1) from (1.12) we get after linearization that

$$\dot{u}(t) - \dot{y}(t) = \frac{\partial f}{\partial y}\Big(t, y(t)\Big)\Big(u(t) - y(t)\Big) + \delta(t) + r(t), \tag{1.13}$$

where, for $t_0 \leq t \leq t_0 + h$, the remainder $r(t)$ is of size $\mathcal{O}\big(\|u(t) - y(t)\|^2\big) = \mathcal{O}(h^{2s+2})$ by Lemma 1.6 below. The variation of constants formula (see e.g., Hairer, Nørsett & Wanner (1993), p. 66) then yields

$$y_1 - y(t_0+h) = u(t_0+h) - y(t_0+h) = \int_{t_0}^{t_0+h} R(t_0+h, s)\Big(\delta(s) + r(s)\Big)\, ds, \tag{1.14}$$

where $R(t, s)$ is the resolvent of the homogeneous part of the differential equation (1.13), i.e., the solution of the matrix differential equation $\partial R(t, s)/\partial t = A(t)R(t, s)$, $R(s, s) = I$, with $A(t) = \partial f/\partial y(t, y(t))$. The integral over $R(t_0 + h, s)r(s)$ gives a $\mathcal{O}(h^{2s+3})$ contribution. The main idea now is to apply the quadrature formula $(b_i, c_i)_{i=1}^s$ to the integral over $g(s) = R(t_0 + h, s)\delta(s)$; because the defect $\delta(s)$ vanishes at the collocation points $t_0 + c_i h$ for $i = 1, \ldots, s$, this gives zero as the numerical result. Thus, the integral is equal to the quadrature error, which is bounded by h^{p+1} times a bound of the pth derivative of the function $g(s)$. This derivative is bounded independently of h, because by Lemma 1.6 all derivatives of the collocation polynomial are bounded uniformly as $h \to 0$. Since, anyway, $p \leq 2s$, we get $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$ from (1.14). □

Lemma 1.6. *The collocation polynomial $u(t)$ is an approximation of order s to the exact solution of (1.1) on the whole interval, i.e.,*

$$\|u(t) - y(t)\| \le C \cdot h^{s+1} \qquad \textit{for} \quad t \in [t_0, t_0 + h] \tag{1.15}$$

and for sufficiently small h.

Moreover, the derivatives of $u(t)$ satisfy for $t \in [t_0, t_0 + h]$

$$\|u^{(k)}(t) - y^{(k)}(t)\| \le C \cdot h^{s+1-k} \qquad \textit{for} \quad k = 0, \ldots, s.$$

Proof. The collocation polynomial satisfies

$$\dot u(t_0 + \tau h) = \sum_{i=1}^{s} f\Big(t_0 + c_i h, u(t_0 + c_i h)\Big)\, \ell_i(\tau),$$

while the exact solution of (1.1) satisfies

$$\dot y(t_0 + \tau h) = \sum_{i=1}^{s} f\Big(t_0 + c_i h, y(t_0 + c_i h)\Big)\, \ell_i(\tau) + h^s E(\tau, h),$$

where the interpolation error $E(\tau, h)$ is bounded by $\max_{t \in [t_0, t_0+h]} \|y^{(s+1)}(t)\|/s!$ and its derivatives satisfy

$$\|E^{(k-1)}(\tau, h)\| \;\le\; \max_{t \in [t_0, t_0+h]} \frac{\|y^{(s+1)}(t)\|}{(s-k+1)!}\,.$$

This follows from the fact that, by Rolle's theorem, the differentiated polynomial $\sum_{i=1}^{s} f\big(t_0 + c_i h, y(t_0 + c_i h)\big)\, \ell_i^{(k-1)}(\tau)$ can be interpreted as the interpolation polynomial of $h^{k-1} y^{(k)}(t_0 + \tau h)$ at $s - k + 1$ points lying in $[t_0, t_0 + h]$. Integrating the difference of the above two equations gives

$$y(t_0 + \tau h) - u(t_0 + \tau h) = h \sum_{i=1}^{s} \Delta f_i \int_0^\tau \ell_i(\sigma)\, d\sigma + h^{s+1} \int_0^\tau E(\sigma, h)\, d\sigma \tag{1.16}$$

with $\Delta f_i = f\big(t_0 + c_i h, y(t_0 + c_i h)\big) - f\big(t_0 + c_i h, u(t_0 + c_i h)\big)$. Using a Lipschitz condition for $f(t, y)$, this relation yields

$$\max_{t \in [t_0, t_0+h]} \|y(t) - u(t)\| \le h\, C\, L \max_{t \in [t_0, t_0+h]} \|y(t) - u(t)\| + \textit{Const} \cdot h^{s+1},$$

implying the statement (1.15) for sufficiently small $h > 0$.

The proof of the second statement follows from

$$h^k \Big(y^{(k)}(t_0 + \tau h) - u^{(k)}(t_0 + \tau h)\Big) = h \sum_{i=1}^{s} \Delta f_i\, \ell_i^{(k-1)}(\tau) + h^{s+1}\, E^{(k-1)}(\tau, h)$$

by using a Lipschitz condition for $f(t, y)$ and the estimate (1.15). $\square$

II.1.3 Gauss and Lobatto Collocation

Gauss Methods. If we take $c_1, \ldots, c_s$ as the zeros of the sth shifted Legendre polynomial

$$\frac{d^s}{dx^s}\Big(x^s(x-1)^s\Big),$$

the interpolatory quadrature formula has order $p = 2s$, and by Theorem 1.5, the Runge–Kutta (or collocation) method based on these nodes has the same order $2s$. For $s = 1$ we obtain the implicit midpoint rule. The Runge–Kutta coefficients for $s = 2$ (the method of Hammer & Hollingsworth 1955) and $s = 3$ are given in Table 1.1. The proof of the order properties for general s was a sensational result of Butcher (1964a). At that time these methods were considered, at least by the editors of *Math. of Comput.*, to be purely academic without any practical value; 5 years later their A-stability was discovered, 12 years later their B-stability, and 25 years later their symplecticity. Thus, of all the papers in issue No. 85 of *Math. of Comput.*, the one most important to us is the one for which publication was the most difficult.

Table 1.1. Gauss methods of order 4 and 6

$$\begin{array}{c|cc}
\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\
\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}$$

$$\begin{array}{c|ccc}
\frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\
\frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array}$$

Radau Methods. Radau quadrature formulas have the highest possible order, $2s-1$, among quadrature formulas with either $c_1 = 0$ or $c_s = 1$. The corresponding collocation methods for $c_s = 1$ are called Radau IIA methods. They play an important role in the integration of stiff differential equations (see Hairer & Wanner (1996), Sect. IV.8). However, they lack both *symmetry* and *symplecticity*, properties that will be the subjects of later chapters in this book.

Lobatto IIIA Methods. Lobatto quadrature formulas have the highest possible order with $c_1 = 0$ and $c_s = 1$. Under these conditions, the nodes must be the zeros of

$$\frac{d^{s-2}}{dx^{s-2}}\Big(x^{s-1}(x-1)^{s-1}\Big) \tag{1.17}$$

and the quadrature order is $p = 2s - 2$. The corresponding collocation methods are called, for historical reasons, Lobatto IIIA methods. For $s = 2$ we have the implicit trapezoidal rule. The coefficients for $s = 3$ and $s = 4$ are given in Table 1.2.

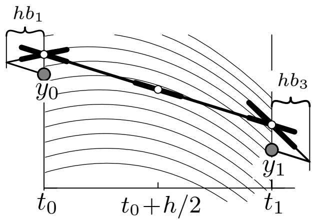
Table 1.2. Lobatto IIIA methods of order 4 and 6

c_i			
0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

c_i				
0	0	0	0	0
$\frac{5-\sqrt{5}}{10}$	$\frac{11+\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	$\frac{-1+\sqrt{5}}{120}$
$\frac{5+\sqrt{5}}{10}$	$\frac{11-\sqrt{5}}{120}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25+\sqrt{5}}{120}$	$\frac{-1-\sqrt{5}}{120}$
1	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

II.1.4 Discontinuous Collocation Methods

Collocation methods allow, as we have seen above, a very elegant proof of their order properties. By similar ideas, they also admit strikingly simple proofs for their A- and B-stability as well as for symplecticity, our subject in Chap. VI. However, not all method classes are of collocation type. It is therefore interesting to define a modification of the collocation idea, which allows us to extend all the above proofs to much wider classes of methods. This definition will also lead, later, to important classes of *partitioned* methods.



Definition 1.7. Let $c_2, \ldots, c_{s-1}$ be distinct real numbers (usually $0 \le c_i \le 1$), and let b_1, b_s be two arbitrary real numbers. The corresponding *discontinuous collocation method* is then defined via a polynomial of degree $s-2$ satisfying

$$\begin{aligned} u(t_0) &= y_0 - hb_1\big(\dot u(t_0) - f(t_0, u(t_0))\big) \\ \dot u(t_0 + c_i h) &= f\big(t_0 + c_i h, u(t_0 + c_i h)\big), \quad i = 2, \ldots, s-1, \\ y_1 &= u(t_1) - hb_s\big(\dot u(t_1) - f(t_1, u(t_1))\big). \end{aligned} \tag{1.18}$$

The figure gives a geometric interpretation of the correction term in the first and third formulas of (1.18). The motivation for this definition will become clear in the proof of Theorem 1.9 below. Our first result shows that discontinuous collocation methods are equivalent to implicit Runge–Kutta methods.

Theorem 1.8. *The discontinuous collocation method of Definition 1.7 is equivalent to an s-stage Runge–Kutta method (1.4) with coefficients determined by $c_1 = 0$, $c_s = 1$, and*

$$\begin{array}{lll} a_{i1} = b_1, & a_{is} = 0 & \text{for } i = 1, \ldots, s, \\ C(s-2) & \text{and} & B(s-2), \end{array} \tag{1.19}$$

with the conditions $C(q)$ and $B(p)$ of (1.11).

Proof. As in the proof of Theorem 1.4 we put $k_i := \dot{u}(t_0 + c_i h)$ (this time for $i = 2, \ldots, s-1$), so that $\dot{u}(t_0+\tau h) = \sum_{j=2}^{s-1} k_j \cdot \ell_j(\tau)$ by the Lagrange interpolation formula. Here, $\ell_j(\tau)$ corresponds to $c_2, \ldots, c_{s-1}$ and is a polynomial of degree $s-3$. By integration and using the definition of $u(t_0)$ we get

$$\begin{aligned} u(t_0 + c_i h) &= u(t_0) + h \sum_{j=2}^{s-1} k_j \int_0^{c_i} \ell_j(\tau)\, d\tau \\ &= y_0 + h b_1 k_1 + h \sum_{j=2}^{s-1} k_j \Big(\int_0^{c_i} \ell_j(\tau)\, d\tau - b_1 \ell_j(0) \Big) \end{aligned}$$

with $k_1 = f(y_0)$. Inserted into (1.18) this gives the first formula of the Runge–Kutta equation (1.4) with $a_{ij} = \int_0^{c_i} \ell_j(\tau)\, d\tau - b_1 \ell_j(0)$. As for collocation methods, one checks that the a_{ij} are uniquely determined by the condition $C(s-2)$. The formula for y_1 is obtained similarly. □

Table 1.3. Survey of discontinuous collocation methods

type	characteristics	prominent examples
$b_1 = 0,\ b_s = 0$	$(s-2)$-stage collocation	Gauss, Radau IIA, Lobatto IIIA
$b_1 = 0,\ b_s \neq 0$	$(s-1)$-stage with $a_{is} = 0$	methods of Butcher (1964b)
$b_1 \neq 0,\ b_s = 0$	$(s-1)$-stage with $a_{i1} = b_1$	Radau IA, Lobatto IIIC
$b_1 \neq 0,\ b_s \neq 0$	s-stage with $a_{i1} = b_1,\ a_{is} = 0$	Lobatto IIIB

If $b_1 = 0$ in Definition 1.7, the entire first column in the Runge–Kutta tableau vanishes, so that the first stage can be removed, which leads to an equivalent method with $s-1$ stages. Similarly, if $b_s = 0$, we can remove the last stage. Therefore, we have all classes of methods, which are "continuous" either to the left, or to the right, or on both sides, as special cases in our definition.

In the case where $b_1 = b_s = 0$, the discontinuous collocation method (1.18) is equivalent to the $(s-2)$-stage collocation method based on $c_2, \ldots, c_{s-1}$ (see Table 1.3). The methods with $b_s = 0$ but $b_1 \neq 0$, which include the Radau IA and

Table 1.4. Lobatto IIIB methods of order 4 and 6

$$
\begin{array}{c|ccc}
0 & \frac{1}{6} & -\frac{1}{6} & 0 \\
\frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 \\
1 & \frac{1}{6} & \frac{5}{6} & 0 \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
\qquad
\begin{array}{c|cccc}
0 & \frac{1}{12} & \frac{-1-\sqrt{5}}{24} & \frac{-1+\sqrt{5}}{24} & 0 \\
\frac{5-\sqrt{5}}{10} & \frac{1}{12} & \frac{25+\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & 0 \\
\frac{5+\sqrt{5}}{10} & \frac{1}{12} & \frac{25+13\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & 0 \\
1 & \frac{1}{12} & \frac{11-\sqrt{5}}{24} & \frac{11+\sqrt{5}}{24} & 0 \\
\hline
 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12}
\end{array}
$$

Lobatto IIIC methods, are of interest for the solution of stiff differential equations (Hairer & Wanner 1996). The methods with $b_1 = 0$ but $b_s \neq 0$, introduced by Butcher (1964a, 1964b), are of historical interest. They were thought to be computationally attractive, because their last stage is explicit. In the context of geometric integration, much more important are methods for which both $b_1 \neq 0$ and $b_s \neq 0$.

Lobatto IIIB Methods (Table 1.4). We consider the quadrature formulas whose nodes are the zeros of (1.17). We have $c_1 = 0$ and $c_s = 1$. Based on $c_2, \ldots, c_{s-1}$ and b_1, b_s we consider the discontinuous collocation method. This class of methods is called Lobatto IIIB (Ehle 1969), and it plays an important role in geometric integration in conjunction with the Lobatto IIIA methods of Sect. II.1.3 (see Theorem IV.2.3 and Theorem VI.4.5). These methods are of order $2s-2$, as the following result shows.

Theorem 1.9 (Superconvergence). *The discontinuous collocation method of Definition 1.7 has the same order as the underlying quadrature formula.*

Proof. We follow the lines of the proof of Theorem 1.5. With the polynomial $u(t)$ of Definition 1.7, and with the defect

$$\delta(t) := \dot{u}(t) - f\big(t, u(t)\big)$$

we get (1.13) after linearization. The variation of constants formula then yields

$$
\begin{aligned}
u(t_0+h) - y(t_0+h) &= R(t_0+h, t_0)\big(u(t_0) - y_0\big) \\
&+ \int_{t_0}^{t_0+h} R(t_0+h, s)\Big(\delta(s) + r(s)\Big)\, ds,
\end{aligned}
$$

which corresponds to (1.14) if $u(t_0) = y_0$. As a consequence of Lemma 1.10 below (with $k = 0$), the integral over $R(t_0+h, s)r(s)$ gives a $\mathcal{O}(h^{2s-1})$ contribution. Since the defect $\delta(t_0 + c_i h)$ vanishes only for $i = 2, \ldots, s-1$, an application of the quadrature formula to $R(t_0+h, s)\delta(s)$ yields $hb_1 R(t_0+h, t_0)\delta(t_0) + hb_s\delta(t_0+h)$ in addition to the quadrature error, which is $\mathcal{O}(h^{p+1})$. Collecting terms suitably, we obtain

$$\begin{aligned} u(t_1) - hb_s\delta(t_1) - y(t_1) &= R(t_1, t_0)\big(u(t_0) + hb_1\delta(t_0) - y_0\big) \\ &\quad + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{2s-1}), \end{aligned}$$

which, after using the definitions of $u(t_0)$ and $u(t_1)$, proves $y_1 - y(t_1) = \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{2s-1})$. □

Lemma 1.10. *The polynomial $u(t)$ of the discontinuous collocation method (1.18) satisfies for $t \in [t_0, t_0 + h]$ and for sufficiently small h*

$$\|u^{(k)}(t) - y^{(k)}(t)\| \le C \cdot h^{s-1-k} \qquad \textit{for} \quad k = 0, \ldots, s-2.$$

Proof. The proof is essentially the same as that for Lemma 1.6. In the formulas for $\dot{u}(t_0 + \tau h)$ and $\dot{y}(t_0 + \tau h)$, the sum has to be taken from $i = 2$ to $i = s - 1$. Moreover, all h^s become h^{s-2}. In (1.16) one has an additional term

$$y_0 - u(t_0) = hb_1\big(\dot{u}(t_0) - f(t_0, u(t_0))\big),$$

which, however, is just an interpolation error of size $\mathcal{O}(h^{s-1})$ and can be included in $Const \cdot h^{s-1}$. □

II.2 Partitioned Runge–Kutta Methods

Some interesting numerical methods introduced in Chap. I (symplectic Euler and the Störmer–Verlet method) do not belong to the class of Runge–Kutta methods. They are important examples of so-called partitioned Runge–Kutta methods. In this section we consider differential equations in the partitioned form

$$\dot{y} = f(y, z), \qquad \dot{z} = g(y, z), \tag{2.1}$$

where y and z may be vectors of different dimensions.

II.2.1 Definition and First Examples

The idea is to take two different Runge–Kutta methods, and to treat the y-variables with the first method (a_{ij}, b_i), and the z-variables with the second method ($\widehat{a}_{ij}, \widehat{b}_i$).

Definition 2.1. Let b_i, a_{ij} and $\widehat{b}_i, \widehat{a}_{ij}$ be the coefficients of two Runge–Kutta methods. A *partitioned Runge–Kutta method* for the solution of (2.1) is given by

$$\begin{aligned} k_i &= f\Big(y_0 + h\sum_{j=1}^{s} a_{ij}k_j,\; z_0 + h\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j\Big), \\ \ell_i &= g\Big(y_0 + h\sum_{j=1}^{s} a_{ij}k_j,\; z_0 + h\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j\Big), \\ y_1 &= y_0 + h\sum_{i=1}^{s} b_i k_i, \qquad z_1 = z_0 + h\sum_{i=1}^{s} \widehat{b}_i \ell_i. \end{aligned} \tag{2.2}$$

Methods of this type were originally proposed by Hofer in 1976 and by Griepentrog in 1978 for problems with stiff and nonstiff parts (see Hairer, Nørsett & Wanner (1993), Sect. II.15). Their importance for Hamiltonian systems (see the examples of Chap. I) has been discovered only in the last decade.

An interesting example is the symplectic Euler method (I.1.9), where the implicit Euler method $b_1 = 1, a_{11} = 1$ is combined with the explicit Euler method $\widehat{b}_1 = 1, \widehat{a}_{11} = 0$. The Störmer–Verlet method (I.1.17) is of the form (2.2) with coefficients given in Table 2.1.

Table 2.1. Störmer–Verlet as a partitioned Runge–Kutta method

c_i	a_{i1}	a_{i2}		$\widehat{c}_i$	$\widehat{a}_{i1}$	$\widehat{a}_{i2}$
0	0	0		1/2	1/2	0
1	1/2	1/2		1/2	1/2	0
	1/2	1/2			1/2	1/2

The theory of Runge–Kutta methods can be extended in a straightforward manner to partitioned methods. Since (2.2) is a one-step method $(y_1, z_1) = \Phi_h(y_0, z_0)$, the Definition 1.2 of the order applies directly. Considering problems $\dot{y} = f(y)$, $\dot{z} = g(z)$ without any coupling terms, we see that the order of (2.2) cannot exceed $\min(p, \widehat{p})$, where p and $\widehat{p}$ are the orders of the two methods.

Conditions for Order Two. Expanding the exact solution of (2.1) and the numerical solution (2.2) into Taylor series, we see that the method is of order 2 if the coupling conditions

$$\textstyle\sum_{ij} b_i \widehat{a}_{ij} = 1/2, \qquad \sum_{ij} \widehat{b}_i a_{ij} = 1/2 \tag{2.3}$$

are satisfied in addition to the usual Runge–Kutta order conditions for order 2. The method of Table 2.1 satisfies these conditions, and it is therefore of order 2. We also remark that (2.3) is automatically satisfied by partitioned methods that are based on the same quadrature nodes, i.e.,

$$c_i = \widehat{c}_i \qquad \text{for all } i \tag{2.4}$$

where, as usual, $c_i = \sum_j a_{ij}$ and $\widehat{c}_i = \sum_j \widehat{a}_{ij}$.

Conditions for Order Three. The conditions for order three already become quite complicated, unless (2.4) is satisfied. In this case, we obtain the additional conditions

$$\textstyle\sum_{ij} b_i \widehat{a}_{ij} c_j = 1/6, \qquad \sum_{ij} \widehat{b}_i a_{ij} c_j = 1/6. \tag{2.5}$$

The order conditions for higher order will be discussed in Sect. III.2.2. It turns out that the number of coupling conditions increases very fast with order, and the proofs for high order are often very cumbersome. There is, however, a very elegant proof of the order for the partitioned method which is the most important one in connection with “geometric integration”, as we shall see now.

II.2.2 Lobatto IIIA–IIIB Pairs

These methods generalize the Störmer–Verlet method to arbitrary order. Indeed, the left method of Table 2.1 is the trapezoidal rule, which is the Lobatto IIIA method with $s = 2$, and the method to the right is equivalent to the midpoint rule and, apart from the values of the c_i, is the Lobatto IIIB method with $s = 2$. Sun (1993b) and Jay (1996) discovered that for general s the combination of the Lobatto IIIA and IIIB methods are suitable for Hamiltonian systems. The coefficients of the methods for $s = 3$ are given in Table 2.2. Using the idea of discontinuous collocation, we give a direct proof of the order for this pair of methods.

Table 2.2. Coefficients of the 3-stage Lobatto IIIA–IIIB pair

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
1/2 & 5/24 & 1/3 & -1/24 \\
1 & 1/6 & 2/3 & 1/6 \\
\hline
 & 1/6 & 2/3 & 1/6
\end{array}
\qquad\qquad
\begin{array}{c|ccc}
0 & 1/6 & -1/6 & 0 \\
1/2 & 1/6 & 1/3 & 0 \\
1 & 1/6 & 5/6 & 0 \\
\hline
 & 1/6 & 2/3 & 1/6
\end{array}
$$

Theorem 2.2. *The partitioned Runge–Kutta method composed of the s-stage Lobatto IIIA and the s-stage Lobatto IIIB method, is of order $2s-2$.*

Proof. Let $c_1 = 0, c_2, \ldots, c_{s-1}, c_s = 1$ and $b_1, \ldots, b_s$ be the nodes and weights of the Lobatto quadrature. The partitioned Runge–Kutta method based on the Lobatto IIIA–IIIB pair can be interpreted as the discontinuous collocation method

$$
\begin{aligned}
&u(t_0) = y_0 \\
&v(t_0) = z_0 - hb_1\big(\dot v(t_0) - g(u(t_0), v(t_0))\big) \\
&\dot u(t_0 + c_i h) = f\big(u(t_0 + c_i h), v(t_0 + c_i h)\big), \qquad i = 1, \ldots, s \\
&\dot v(t_0 + c_i h) = g\big(u(t_0 + c_i h), v(t_0 + c_i h)\big), \qquad i = 2, \ldots, s-1 \\
&y_1 = u(t_1) \\
&z_1 = v(t_1) - hb_s\big(\dot v(t_1) - g(u(t_1), v(t_1))\big),
\end{aligned}
\tag{2.6}
$$

where $u(t)$ and $v(t)$ are polynomials of degree s and $s-2$, respectively. This is seen as in the proofs of Theorem 1.4 and Theorem 1.8. The superconvergence (order $2s-2$) is obtained with exactly the same proof as for Theorem 1.9, where the functions $u(t)$ and $y(t)$ have to be replaced with $(u(t), v(t))^T$ and $(y(t), z(t))^T$, etc. Instead of Lemma 1.10 we use the estimates (for $t \in [t_0, t_0 + h]$)

$$
\begin{aligned}
\|u^{(k)}(t) - y^{(k)}(t)\| &\le c \cdot h^{s-k} \quad \text{for} \quad k = 0, \ldots, s, \\
\|v^{(k)}(t) - z^{(k)}(t)\| &\le c \cdot h^{s-1-k} \quad \text{for} \quad k = 0, \ldots, s-2,
\end{aligned}
$$

which can be proved by following the lines of the proofs of Lemma 1.6 and Lemma 1.10. □

II.2.3 Nyström Methods

> Da bis jetzt die *direkte* Anwendung der Rungeschen Methode auf den wichtigen Fall von Differentialgleichungen zweiter Ordnung nicht behandelt war . . . (E.J. Nyström 1925)

Second-order differential equations

$$\ddot{y} = g(t, y, \dot{y}) \tag{2.7}$$

form an important class of problems. Most of the differential equations in Chap. I are of this form (e.g., the Kepler problem, the outer solar system, problems in molecular dynamics). This is mainly due to Newton's law that forces are proportional to second derivatives (acceleration). Introducing a new variable $z = \dot{y}$ for the first derivative, the problem (2.7) becomes equivalent to the partitioned system

$$\dot{y} = z, \qquad \dot{z} = g(t, y, z). \tag{2.8}$$

A partitioned Runge–Kutta method (2.2) applied to this system yields

$$\begin{aligned} k_i &= z_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} \ell_j, \\ \ell_i &= g\Big(t_0 + c_i h,\; y_0 + h \sum_{j=1}^{s} a_{ij} k_j,\; z_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} \ell_j\Big), \\ y_1 &= y_0 + h \sum_{i=1}^{s} b_i k_i, \qquad z_1 \;=\; z_0 + h \sum_{i=1}^{s} \widehat{b}_i \ell_i. \end{aligned} \tag{2.9}$$

If we insert the formula for k_i into the others, we obtain Definition 2.3 with

$$\overline{a}_{ij} = \sum_{k=1}^{s} a_{ik} \widehat{a}_{kj}, \qquad \overline{b}_i = \sum_{k=1}^{s} b_k \widehat{a}_{ki}. \tag{2.10}$$

Definition 2.3. Let $c_i, \overline{b}_i, \overline{a}_{ij}$ and $\widehat{b}_i, \widehat{a}_{ij}$ be real coefficients. A *Nyström method* for the solution of (2.7) is given by

$$\begin{aligned} \ell_i &= g\Big(t_0 + c_i h,\; y_0 + c_i h \dot{y}_0 + h^2 \sum_{j=1}^{s} \overline{a}_{ij} \ell_j,\; \dot{y}_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} \ell_j\Big), \\ y_1 &= y_0 + h \dot{y}_0 + h^2 \sum_{i=1}^{s} \overline{b}_i \ell_i, \qquad \dot{y}_1 \;=\; \dot{y}_0 + h \sum_{i=1}^{s} \widehat{b}_i \ell_i. \end{aligned} \tag{2.11}$$

For the important special case $\ddot{y} = g(t, y)$, where the vector field does not depend on the velocity, the coefficients $\widehat{a}_{ij}$ need not be specified. A Nyström method is of order p if $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$ and $\dot{y}_1 - \dot{y}(t_0 + h) = \mathcal{O}(h^{p+1})$. It is not sufficient to consider y_1 alone. The order conditions will be discussed in Sect. III.2.3.

Notice that the Störmer–Verlet scheme (I.1.17) is a Nyström method for problems of the form $\ddot{y} = g(t, y)$. We have $s = 2$, and the coefficients are $c_1 = 0, c_2 = 1$, $\overline{a}_{11} = \overline{a}_{12} = \overline{a}_{22} = 0$, $\overline{a}_{21} = 1/2$, $\overline{b}_1 = 1/2$, $\overline{b}_2 = 0$, and $\widehat{b}_1 = \widehat{b}_2 = 1/2$. With $q_{n+1/2} = q_n + \frac{h}{2} v_{n+1/2}$ the step $(q_{n-1/2}, v_{n-1/2}) \mapsto (q_{n+1/2}, v_{n+1/2})$ of (I.1.17) becomes a one-stage Nyström method with $c_1 = 1/2$, $\overline{a}_{11} = 0$, $\overline{b}_1 = \widehat{b}_1 = 1$.

II.3 The Adjoint of a Method

We shall see in Chap. V that *symmetric* numerical methods have many important properties. The key for understanding symmetry is the concept of the *adjoint* method.

The flow φ_t of an autonomous differential equation

$$\dot{y} = f(y), \qquad y(t_0) = y_0 \tag{3.1}$$

satisfies $\varphi_{-t}^{-1} = \varphi_t$. This property is *not*, in general, shared by the one-step map Φ_h of a numerical method. An illustration is presented in the upper picture of Fig. 3.1 (a), where we see that the one-step map Φ_h for the explicit Euler method is different from the inverse of Φ_{-h}, which is the implicit Euler method.

Definition 3.1. The *adjoint method* Φ_h^* of a method Φ_h is the inverse map of the original method with reversed time step $-h$, i.e.,

$$\Phi_h^* := \Phi_{-h}^{-1} \tag{3.2}$$

(see Fig. 3.1 (b)). In other words, $y_1 = \Phi_h^*(y_0)$ is implicitly defined by $\Phi_{-h}(y_1) = y_0$. A method for which $\Phi_h^* = \Phi_h$ is called *symmetric*.

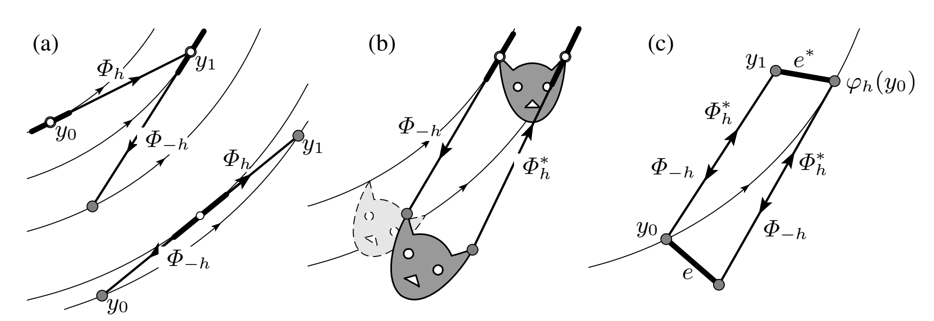


Fig. 3.1. Definition and properties of the adjoint method

The consideration of adjoint methods evolved independently from the study of symmetric integrators (Stetter (1973), p. 125, Wanner (1973)) and from the aim of constructing and analyzing stiff integrators from explicit ones (Cash (1975) calls them "the backward version" which were the first example of mono-implicit methods and Scherer (1977) calls them "reflected methods").

The adjoint method satisfies the usual properties such as $(\Phi_h^*)^* = \Phi_h$ and $(\Phi_h \circ \Psi_h)^* = \Psi_h^* \circ \Phi_h^*$ for any two one-step methods Φ_h and Ψ_h. The implicit Euler method is the adjoint of the explicit Euler method. The implicit midpoint rule is symmetric (see the lower picture of Fig. 3.1 (a)), and the trapezoidal rule and the Störmer–Verlet method are also symmetric.

The following theorem shows that the adjoint method has the same order as the original method, and, with a possible sign change, also the same leading error term.

Theorem 3.2. *Let φ_t be the exact flow of (3.1) and let Φ_h be a one-step method of order p satisfying*

$$\Phi_h(y_0) = \varphi_h(y_0) + C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}). \tag{3.3}$$

The adjoint method Φ_h^* then has the same order p and we have

$$\Phi_h^*(y_0) = \varphi_h(y_0) + (-1)^p C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}). \tag{3.4}$$

If the method is symmetric, its (maximal) order is even.

Proof. The idea of the proof is exhibited in drawing (c) of Fig. 3.1. From a given initial value y_0 we compute $\varphi_h(y_0)$ and $y_1 = \Phi_h^*(y_0)$, whose difference e^* is the local error of Φ_h^*. This error is then "projected back" by Φ_{-h} to become e. We see that $-e$ is the local error of Φ_{-h}, i.e., by hypothesis (3.3),

$$e = (-1)^p C(\varphi_h(y_0))h^{p+1} + \mathcal{O}(h^{p+2}). \tag{3.5}$$

Since $\varphi_h(y_0) = y_0 + \mathcal{O}(h)$ and $e = (I + \mathcal{O}(h))e^*$, it follows that

$$e^* = (-1)^p C(y_0)h^{p+1} + \mathcal{O}(h^{p+2})$$

which proves (3.4). The statement for symmetric methods is an immediate consequence of this result, because $\Phi_h = \Phi_h^*$ implies $C(y_0) = (-1)^p C(y_0)$, and therefore $C(y_0)$ can be different from zero only for even p. □

II.4 Composition Methods

The idea of composing methods has some tradition in several variants: composition of different Runge–Kutta methods with the same step size leading to the Butcher group, which is treated in Sect. III.1.3; cyclic composition of multistep methods for breaking the "Dahlquist barrier" (see Stetter (1973), p. 216); composition of low order Runge–Kutta methods for increasing stability for stiff problems (Gentzsch & Schlüter (1978), Iserles (1984)). In the following, we consider the composition of a given basic one-step method (and, eventually, its adjoint method) with *different* step sizes. The aim is to increase the order while preserving some desirable properties of the basic method. This idea has mainly been developed in the papers of Suzuki (1990), Yoshida (1990), and McLachlan (1995).

Let Φ_h be a basic method and $\gamma_1, \ldots, \gamma_s$ real numbers. Then we call its composition with step sizes $\gamma_1 h, \gamma_2 h, \ldots, \gamma_s h$, i.e.,

$$\Psi_h = \Phi_{\gamma_s h} \circ \ldots \circ \Phi_{\gamma_1 h}, \tag{4.1}$$

the corresponding *composition method* (see Fig. 4.1 (a)).

Theorem 4.1. *Let Φ_h be a one-step method of order p. If*

$$\begin{aligned} \gamma_1 + \ldots + \gamma_s &= 1 \\ \gamma_1^{p+1} + \ldots + \gamma_s^{p+1} &= 0, \end{aligned} \tag{4.2}$$

then the composition method (4.1) is at least of order $p+1$.

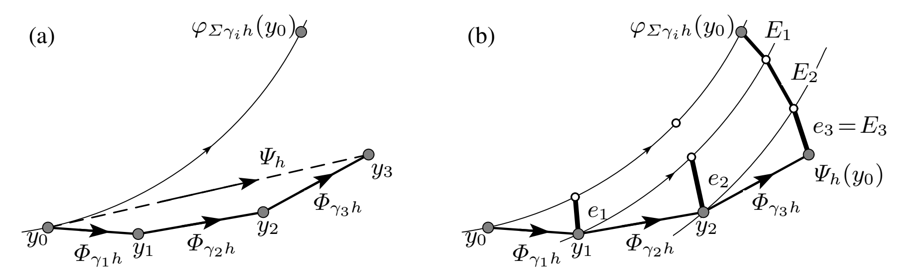


Fig. 4.1. Composition of method Φ_h with three step sizes

Proof. The proof is presented in Fig. 4.1 (b) for $s = 3$. It is very similar to the proof of Theorem 3.2. By hypothesis

$$\begin{aligned} e_1 &= C(y_0) \cdot \gamma_1^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}) \\ e_2 &= C(y_1) \cdot \gamma_2^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}) \\ e_3 &= C(y_2) \cdot \gamma_3^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}). \end{aligned} \tag{4.3}$$

We have, as before, $y_i = y_0 + \mathcal{O}(h)$ and $E_i = (I + \mathcal{O}(h))e_i$ for all i and obtain, for $\sum \gamma_i = 1$,

$$\varphi_h(y_0) - \Psi_h(y_0) = E_1 + E_2 + E_3 = C(y_0)(\gamma_1^{p+1} + \gamma_2^{p+1} + \gamma_3^{p+1})h^{p+1} + \mathcal{O}(h^{p+2})$$

which shows that under conditions (4.2) the $\mathcal{O}(h^{p+1})$-term vanishes. □

Example 4.2 (The Triple Jump). Equations (4.2) have no real solution for odd p. Therefore, the order increase is only possible for even p. In this case, the smallest s which allows a solution is $s = 3$. We then have some freedom for solving the two equations. If we impose symmetry $\gamma_1 = \gamma_3$, then we obtain (Creutz & Gocksch 1989, Forest 1989, Suzuki 1990, Yoshida 1990)

$$\gamma_1 = \gamma_3 = \frac{1}{2 - 2^{1/(p+1)}}, \qquad \gamma_2 = -\frac{2^{1/(p+1)}}{2 - 2^{1/(p+1)}}\,. \tag{4.4}$$

This procedure can be repeated: we start with a symmetric method of order 2, apply (4.4) with $p = 2$ to obtain order 3; due to the symmetry of the γ's this new method is in fact of order 4 (see Theorem 3.2). With this new method we repeat (4.4) with $p = 4$ and obtain a symmetric 9-stage composition method of order 6, then with $p = 6$ a 27-stage symmetric composition method of order 8, and so on. One obtains in this way *any* order, however, at the price of a terrible zig-zag of the step points (see Fig. 4.2).

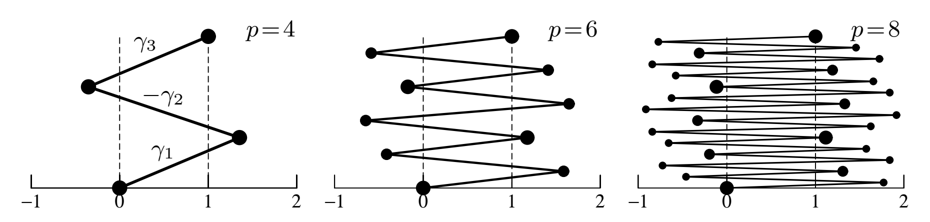


Fig. 4.2. The Triple Jump of order 4 and its iterates of orders 6 and 8

Example 4.3 (Suzuki's Fractals). If one desires methods with smaller values of γ_i, one has to increase s even more. For example, for $s = 5$ the best solution of (4.2) has the sign structure $+ + - + +$ with $\gamma_1 = \gamma_2$ (see Exercise 7). This leads to (Suzuki 1990)

$$\gamma_1 = \gamma_2 = \gamma_4 = \gamma_5 = \frac{1}{4 - 4^{1/(p+1)}}, \qquad \gamma_3 = -\frac{4^{1/(p+1)}}{4 - 4^{1/(p+1)}} . \tag{4.5}$$

The repetition of this algorithm for $p = 2, 4, 6, \ldots$ leads to a fractal structure of the step points (see Fig. 4.3).

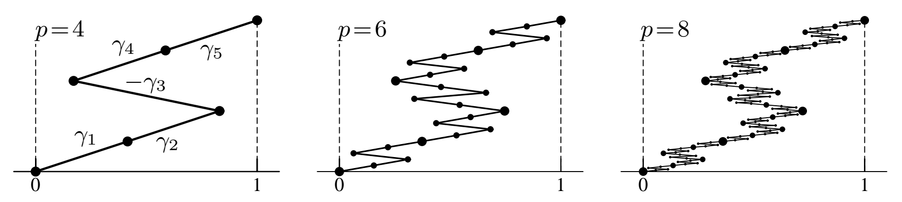


Fig. 4.3. Suzuki's "fractal" composition methods

Composition with the Adjoint Method. If we replace the composition (4.1) by the more general formula

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi^*_{\beta_s h} \circ \ldots \circ \Phi^*_{\beta_2 h} \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}, \tag{4.6}$$

the condition for order $p + 1$ becomes, by using the result (3.4) and a similar proof as above,

$$\begin{aligned} \beta_1 + \alpha_1 + \beta_2 + \ldots + \beta_s + \alpha_s &= 1 \\ (-1)^p \beta_1^{p+1} + \alpha_1^{p+1} + (-1)^p \beta_2^{p+1} + \ldots + (-1)^p \beta_s^{p+1} + \alpha_s^{p+1} &= 0. \end{aligned} \tag{4.7}$$

This allows an order increase for odd p as well. In particular, we see at once the solution $\alpha_1 = \beta_1 = 1/2$ for $p = s = 1$, which turns every consistent one-step method of order 1 into a second-order symmetric method

$$\Psi_h = \Phi_{h/2} \circ \Phi^*_{h/2}. \tag{4.8}$$

Example 4.4. If Φ_h is the explicit (resp. implicit) Euler method, then Ψ_h in (4.8) becomes the implicit midpoint (resp. trapezoidal) rule.

Example 4.5. In a second-order problem $\dot{q} = p$, $\dot{p} = g(q)$, if Φ_h is the symplectic Euler method, which discretizes q by the implicit Euler and p by the explicit Euler method, then the composed method Ψ_h in (4.8) is the Störmer–Verlet method (I.1.17).

A Numerical Example. To demonstrate the numerical performance of the above methods, we choose the Kepler problem (I.2.2) with $e = 0.6$ and the initial values from (I.2.11). As integration interval we choose $[0, 7.5]$, a bit more than one revolution. The exact solution is obtained by carefully evaluating the integral (I.2.10), which gives

$$\varphi = 8.670026323142814951591088285 52, \tag{4.9}$$

with the help of which we compute r, $\dot\varphi$, $\dot r$ from (I.2.8) and (I.2.6). This gives

$$\begin{aligned} q_1 &= -0.828164402690770818204757585370 \\ q_2 &= 0.778898095658635447081654480796 \\ p_1 &= -0.856384715343395351524486215030 \\ p_2 &= -0.160552150799838435254419104102\,. \end{aligned} \tag{4.10}$$

As the basic method we use the Verlet scheme and compare in Fig. 4.4 the performances of the composition sequences of the Triple Jump (4.4) and those of Suzuki (4.5) for a large number of different equidistant basic step sizes and for orders $p = 4, 6, 8, 10, 12$. Each basic step is then divided into $3, 9, 27, 81, 243$ respectively $5, 25, 125, 625, 3125$ composition steps and the maximal final error is compared with the total number of function evaluations in double logarithmic scales. For each method and order, all the points lie asymptotically on a straight line with slope $-p$. Therefore, theoretically, a higher order method will become superior when the precision requirements become sufficiently high. But we see that for orders 10 and 12 these "break even points" are far beyond any precision of practical interest, after some 40 or 50 digits. We also observe that the wild zig-zag of the Triple Jump (4.4) is a more serious handicap than the enormous number of small steps of the Suzuki sequence (4.5).

For later reference we have also included, in black symbols, the results obtained by the two methods (V.3.11) and (V.3.13) of orders 6 and 8, respectively, which will be the outcome of a more elaborate order theory of Chap. III.

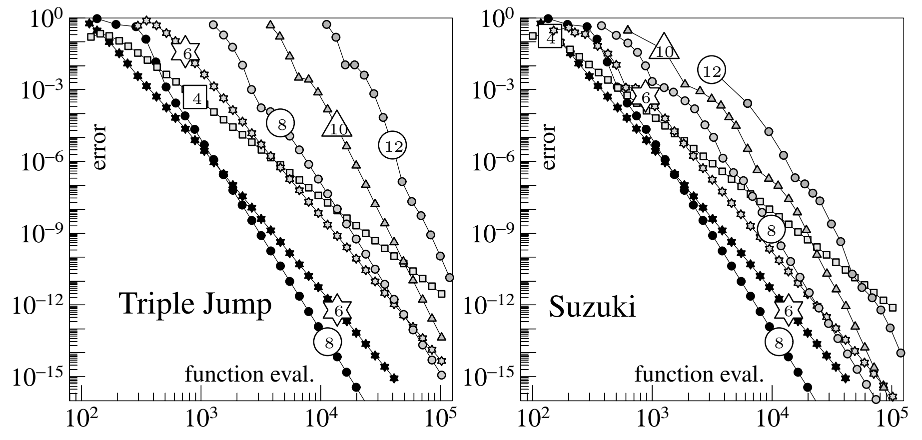


Fig. 4.4. Numerical results of the Triple Jump and Suzuki step sequences (grey symbols) compared to optimal methods (black symbols)

II.5 Splitting Methods

The splitting idea yields an approach that is completely different from Runge–Kutta methods. One decomposes the vector field into integrable pieces and treats them separately.

Fig. 5.1. A splitting of a vector field

We consider an arbitrary system $\dot{y} = f(y)$ in $\mathbb{R}^n$, and suppose that the vector field is "split" as (see Fig. 5.1)

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y). \tag{5.1}$$

If then, by chance, the exact flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems $\dot{y} = f^{[1]}(y)$ and $\dot{y} = f^{[2]}(y)$ can be calculated explicitly, we can, from a given initial value y_0, first solve the first system to obtain a value $y_{1/2}$, and from this value integrate the second system to obtain y_1. In this way we have introduced the numerical methods

$$\Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]} \qquad \Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \tag{5.2}$$

where one is the adjoint of the other. These formulas are often called the *Lie–Trotter splitting* (Trotter 1959). By Taylor expansion we find that $(\varphi_h^{[1]} \circ \varphi_h^{[2]})(y_0) = \varphi_h(y_0) + \mathcal{O}(h^2)$, so that both methods give approximations of order 1 to the solution of (5.1). Another idea is to use a symmetric version and put

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}, \tag{5.3}$$

which is known as the *Strang splitting*[1] (Strang 1968), and sometimes as the *Marchuk splitting* (Marchuk 1968). By breaking up in (5.3) $\varphi_h^{[2]} = \varphi_{h/2}^{[2]} \circ \varphi_{h/2}^{[2]}$,

[1] The article Strang (1968) deals with spatial discretizations of partial differential equations such as $u_t = Au_x + Bu_y$. There, the functions $f^{[i]}$ typically contain differences in only one spatial direction.

we see that the Strang splitting $\Phi_h^{[S]} = \Phi_{h/2} \circ \Phi_{h/2}^*$ is the composition of the Lie-Trotter method and its adjoint with halved step sizes. The Strang splitting formula is therefore symmetric and of order 2 (see (4.8)).

Example 5.1 (The Symplectic Euler and the Störmer–Verlet Schemes). Suppose we have a Hamiltonian system with separable Hamiltonian $H(p,q) = T(p) + U(q)$. We consider this as the sum of *two* Hamiltonians, the first one depending only on p, the second one only on q. The corresponding Hamiltonian systems

$$\begin{aligned} \dot p &= 0 \\ \dot q &= T_p(p) \end{aligned} \qquad \text{and} \qquad \begin{aligned} \dot p &= -U_q(q) \\ \dot q &= 0 \end{aligned} \tag{5.4}$$

can be solved without problem to yield

$$\begin{aligned} p(t) &= p_0 \\ q(t) &= q_0 + t\,T_p(p_0) \end{aligned} \qquad \text{and} \qquad \begin{aligned} p(t) &= p_0 - t\,U_q(q_0) \\ q(t) &= q_0. \end{aligned} \tag{5.5}$$

Denoting the flows of these two systems by φ_t^T and φ_t^U, we see that the symplectic Euler method (I.1.9) is just the composition $\varphi_h^T \circ \varphi_h^U$. Furthermore, the adjoint of the symplectic Euler method is $\varphi_h^U \circ \varphi_h^T$, and by Example 4.5 the Verlet scheme is $\varphi_{h/2}^U \circ \varphi_h^T \circ \varphi_{h/2}^U$, the Strang splitting (5.3). Anticipating the results of Chap. VI, the flows φ_h^T and φ_h^U are both symplectic transformations, and, since the composition of symplectic maps is again symplectic, this gives an elegant proof of the symplecticity of the "symplectic" Euler method and the Verlet scheme.

General Splitting Procedure. In a similar way to the general idea of composition methods (4.6), we can form with arbitrary coefficients $a_1, b_1, a_2, \ldots, a_m, b_m$ (where, eventually, a_1 or b_m, or both, are zero)

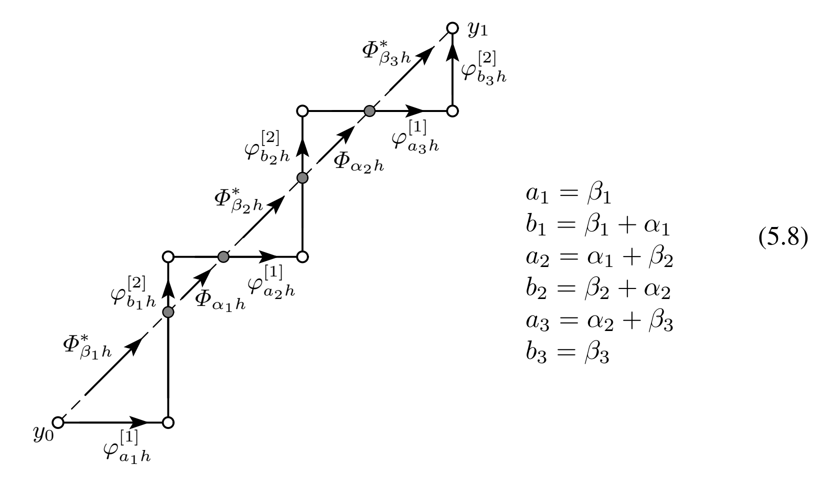
$$\Psi_h = \varphi_{b_m h}^{[2]} \circ \varphi_{a_m h}^{[1]} \circ \varphi_{b_{m-1} h}^{[2]} \circ \ldots \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \tag{5.6}$$

and try to increase the order of the scheme by suitably determining the free coefficients. An early contribution to this subject is the article of Ruth (1983), where, for the special case (5.4), a method (5.6) of order 3 with $m = 3$ is constructed. Forest & Ruth (1990) and Candy & Rozmus (1991) extend Ruth's technique and construct methods of order 4. One of their methods is just (4.1) with $\gamma_1, \gamma_2, \gamma_3$ given by (4.4) ($p = 2$) and Φ_h from (5.3). A systematic study of such methods started with the articles of Suzuki (1990, 1992) and Yoshida (1990).

A close connection between the theories of splitting methods (5.6) and of composition methods (4.6) was discovered by McLachlan (1995). Indeed, if we put $\beta_1 = a_1$ and break up $\varphi_{b_1 h}^{[2]} = \varphi_{\alpha_1 h}^{[2]} \circ \varphi_{\beta_1 h}^{[2]}$ (group property of the exact flow) where α_1 is given in (5.8), further $\varphi_{a_2 h}^{[1]} = \varphi_{\beta_2 h}^{[1]} \circ \varphi_{\alpha_1 h}^{[1]}$ and so on (cf. Fig. 5.2), we see, using (5.2), that Ψ_h of (5.6) is identical with Ψ_h of (4.6), where

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \qquad \text{so that} \qquad \Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]}. \tag{5.7}$$

A necessary and sufficient condition for the existence of α_i and β_i satisfying (5.8) is that $\sum a_i = \sum b_i$, which is the consistency condition anyway for method (5.6).



$$\begin{aligned} a_1 &= \beta_1 \\ b_1 &= \beta_1 + \alpha_1 \\ a_2 &= \alpha_1 + \beta_2 \\ b_2 &= \beta_2 + \alpha_2 \\ a_3 &= \alpha_2 + \beta_3 \\ b_3 &= \beta_3 \end{aligned} \tag{5.8}$$

Fig. 5.2. Equivalence of splitting and composition methods

Combining Exact and Numerical Flows. It may happen that the differential equation $\dot{y} = f(y)$ can be split according to (5.1), such that only the flow of, say, $\dot{y} = f^{[1]}(y)$ can be computed exactly. If $f^{[1]}(y)$ constitutes the dominant part of the vector field, it is natural to search for integrators that exploit this information. The above interpretation of splitting methods as composition methods allows us to construct such integrators. We just consider

$$\Phi_h = \varphi_h^{[1]} \circ \Phi_h^{[2]}, \qquad \Phi_h^* = \Phi_h^{[2]*} \circ \varphi_h^{[1]} \tag{5.9}$$

as the basis of the composition method (4.6). Here $\varphi_t^{[1]}$ is the exact flow of $\dot{y} = f^{[1]}(y)$, and $\Phi_h^{[2]}$ is some first-order integrator applied to $\dot{y} = f^{[2]}(y)$. Since Φ_h of (5.9) is consistent with (5.1), the resulting method (4.6) has the desired high order. It is given by

$$\Psi_h = \varphi^{[1]}_{\alpha_s h} \circ \Phi^{[2]}_{\alpha_s h} \circ \Phi^{[2]*}_{\beta_s h} \circ \varphi^{[1]}_{(\beta_s + \alpha_{s-1})h} \circ \Phi^{[2]}_{\alpha_{s-1} h} \circ \ldots \circ \Phi^{[2]*}_{\beta_1 h} \circ \varphi^{[1]}_{\beta_1 h}. \tag{5.10}$$

Notice that replacing $\varphi_t^{[2]}$ with a low-order approximation $\Phi_t^{[2]}$ in (5.6) would not retain the high order of the composition, because $\Phi_t^{[2]}$ does not satisfy the group property.

Splitting into More than Two Vector Fields. Consider a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y) + \ldots + f^{[N]}(y), \tag{5.11}$$

where we assume that the flows $\varphi_t^{[j]}$ of the individual problems $\dot{y} = f^{[j]}(y)$ can be computed exactly. In this case there are many possibilities for extending (5.6) and for writing the method as a composition of $\varphi^{[1]}_{a_j h}, \varphi^{[2]}_{b_j h}, \varphi^{[3]}_{c_j h}, \ldots$. This makes it difficult to find optimal compositions of high order. A simple and efficient way is to consider the first-order method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \ldots \circ \varphi_h^{[N]}$$

together with its adjoint as the basis of the composition (4.6). Without any additional effort this yields splitting methods for (5.11) of arbitrary high order.

II.6 Exercises

1. Compute all collocation methods with $s = 2$ as a function of c_1 and c_2. Which of them are of order 3, which of order 4?
2. Prove that the collocation solution plotted in the right picture of Fig. 1.3 is composed of arcs of parabolas.
3. Let $b_1 = b_4 = 1/8$, $c_2 = 1/3$, $c_3 = 2/3$, and consider the corresponding discontinuous collocation method. Determine its order and find the coefficients of the equivalent Runge–Kutta method.
4. Show that each of the symplectic Euler methods in (I.1.9) is the adjoint of the other.
5. (Additive Runge–Kutta methods). Let b_i, a_{ij} and $b_i, \widehat{a}_{ij}$ be the coefficients of two Runge–Kutta methods. An additive Runge–Kutta method for the solution of $\dot y = f^{[1]}(y) + f^{[2]}(y)$ is given by

$$\begin{aligned} k_i &= f^{[1]}\Big(y_0 + h\sum_{j=1}^s a_{ij}k_j\Big) + f^{[2]}\Big(y_0 + h\sum_{j=1}^s \widehat{a}_{ij}k_j\Big) \\ y_1 &= y_0 + h\sum_{i=1}^s b_i k_i . \end{aligned}$$

   Show that this can be interpreted as a partitioned Runge–Kutta method (2.2) applied to

$$\dot y = f^{[1]}(y) + f^{[2]}(z), \qquad \dot z = f^{[1]}(y) + f^{[2]}(z)$$

   with $y(0) = z(0) = y_0$. Notice that $y(t) = z(t)$.
6. Let Φ_h denote the Störmer–Verlet scheme, and consider the composition

$$\Phi_{\gamma_{2k+1}h} \circ \Phi_{\gamma_{2k}h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}$$

   with $\gamma_1 = \ldots = \gamma_k = \gamma_{k+2} = \ldots = \gamma_{2k+1}$. Compute γ_1 and γ_{k+1} such that the composition gives a method of order 4. For several differential equations (pendulum, Kepler problem) study the global error of a constant step size implementation as a function of k.
7. Consider the composition method (4.1) with $s = 5$, $\gamma_5 = \gamma_1$, and $\gamma_4 = \gamma_2$. Among the solutions of

$$2\gamma_1 + 2\gamma_2 + \gamma_3 = 1, \qquad 2\gamma_1^3 + 2\gamma_2^3 + \gamma_3^3 = 0$$

   find the one that minimizes $|2\gamma_1^5 + 2\gamma_2^5 + \gamma_3^5|$.
   Remark. This property motivates the choice of the γ_i in (4.5).

Chapter III. Order Conditions, Trees and B-Series

In this chapter we present a compact theory of the order conditions of the methods presented in Chap. II, in particular Runge–Kutta methods, partitioned Runge–Kutta methods, and composition methods by using the notion of rooted trees and B-series. These ideas lead to algebraic structures which have recently found interesting applications in quantum field theory. The chapter terminates with the Baker-Campbell-Hausdorff formula, which allows another access to the order properties of composition and splitting methods.

Some parts of this chapter are rather short, but nevertheless self-contained. For more detailed presentations we refer to the monographs of Butcher (1987), of Hairer, Nørsett & Wanner (1993), and of Hairer & Wanner (1996). Readers mainly interested in geometric properties of numerical integrators may continue with Chapters IV, V or VI before returning to the technically more difficult jungle of trees.

III.1 Runge–Kutta Order Conditions and B-Series

> Even the standard notation has been found to be too heavy in dealing with fourth and higher order processes, . . . (R.H. Merson 1957)

In this section we derive the order conditions of Runge–Kutta methods by comparing the Taylor series of the exact solution of (1.1) with that of the numerical solution. The computation is much simplified, first by considering an *autonomous* system of equations (Gill 1951), and second, by the use of rooted trees (connected graphs without cycles and a distinguished vertex; Merson 1957). The theory has been developed by Butcher in the years 1963-72 (see Butcher (1987), Sect. 30) and by Hairer & Wanner in 1973-74 (see Hairer, Nørsett & Wanner (1993), Sections II.2 and II.12). Here we give new simplified proofs.

III.1.1 Derivation of the Order Conditions

We consider an autonomous problem

$$\dot{y} = f(y), \qquad y(t_0) = y_0, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is sufficiently differentiable. A problem $\dot{y} = f(t, y)$ can be brought into this form by appending the equation $\dot{t} = 1$. We develop the subsequent theory in four steps.

Er sagte es klar und angenehm,
was erstens, zweitens und drittens käm'. (W. Busch, *Jobsiade* 1872)

First Step. We compute the higher derivatives of the solution y at the initial point t_0. For this, we have from (1.1)

$$y^{(q)} = \big(f(y)\big)^{(q-1)} \tag{1.2}$$

and compute the latter derivatives by using the chain rule, the product rule, the symmetry of partial derivatives, and the notation $f'(y)$ for the derivative as a linear map (the Jacobian), $f''(y)$ the second derivative as a bilinear map and similarly for higher derivatives. This gives

$$\begin{aligned}
\dot y &= f(y) \\
\ddot y &= f'(y)\,\dot y \qquad\qquad (1.3)\\
y^{(3)} &= f''(y)(\dot y,\dot y) + f'(y)\,\ddot y \\
y^{(4)} &= f'''(y)(\dot y,\dot y,\dot y) + 3f''(y)(\ddot y,\dot y) + f'(y)\,y^{(3)} \\
y^{(5)} &= f^{(4)}(y)(\dot y,\dot y,\dot y,\dot y) + 6f'''(y)(\ddot y,\dot y,\dot y) + 4f''(y)(y^{(3)},\dot y) \\
&\quad +3f''(y)(\ddot y,\ddot y) + f'(y)\,y^{(4)},
\end{aligned}$$

and so on. The coefficients $3, 6, 4, 3, \ldots$ appearing in these expressions have a certain combinatorial meaning (number of partitions of a set of $q-1$ elements), but for the moment we need not know their values.

Second Step. We insert in (1.3) recursively the computed derivatives $\dot y$, $\ddot y, \ldots$ into the right side of the subsequent formulas. This gives for the first few

$$\begin{aligned}
\dot y &= f \\
\ddot y &= f'f \qquad\qquad (1.4)\\
y^{(3)} &= f''(f,f) + f'f'f \\
y^{(4)} &= f'''(f,f,f) + 3f''(f'f,f) + f'f''(f,f) + f'f'f'f,
\end{aligned}$$

where the arguments (y) have been suppressed. The expressions which appear in these formulas, denoted by $F(\tau)$, will be called the *elementary differentials*. We represent each of them by a suitable graph τ (a rooted tree) as follows:

Each f becomes a vertex, a first derivative f' becomes a vertex with one branch, and a kth derivative $f^{(k)}$ becomes a vertex with k branches pointing upwards. The arguments of the k-linear mapping $f^{(k)}(y)$ correspond to trees that are attached on the upper ends of these branches. The tree to the right corresponds to $f''(f'f,f)$. Other trees are plotted in Table 1.1. In the above process, each insertion of an already known derivative consists of grafting the corresponding trees upon a new root as in Definition 1.1 below, and inserting the corresponding elementary differentials as arguments of $f^{(m)}(y)$ as in Definition 1.2.

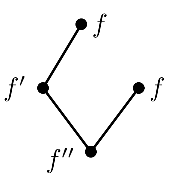


Table 1.1. Trees, elementary differentials, and coefficients

$\|\tau\|$	τ	graph	$\alpha(\tau)$	$F(\tau)$	$\gamma(\tau)$	$\phi(\tau)$	$\sigma(\tau)$
1	$\bullet$	$\bullet$	1	f	1	$\sum_i b_i$	1
2	$[\bullet]$		1	$f'f$	2	$\sum_{ij} b_i a_{ij}$	1
3	$[\bullet,\bullet]$		1	$f''(f,f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$	2
3	$[[\bullet]]$		1	$f'f'f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$	1
4	$[\bullet,\bullet,\bullet]$		1	$f'''(f,f,f)$	4	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{il}$	6
4	$[[\bullet],\bullet]$		3	$f''(f'f,f)$	8	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{jl}$	1
4	$[[\bullet,\bullet]]$		1	$f'f''(f,f)$	12	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{jl}$	2
4	$[[[\bullet]]]$		1	$f'f'f'f$	24	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{kl}$	1

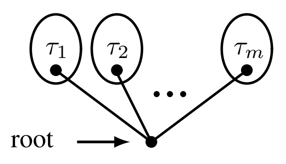
Definition 1.1 (Trees). The set of (rooted) *trees* T is recursively defined as follows:

a) the graph $\bullet$ with only one vertex (called the root) belongs to T;

b) if $\tau_1,\ldots,\tau_m \in T$, then the graph obtained by grafting the roots of $\tau_1,\ldots,\tau_m$ to a new vertex also belongs to T. It is denoted by

$$\tau = [\tau_1,\ldots,\tau_m],$$

and the new vertex is the root of τ.



We further denote by $|\tau|$ the *order* of τ (the number of vertices), and by $\alpha(\tau)$ the coefficients appearing in the formulas (1.4). We remark that some of the trees among $\tau_1,\ldots,\tau_m$ may be equal and that τ does not depend on the ordering of $\tau_1,\ldots,\tau_m$. For example, we do not distinguish between $[[\bullet],\bullet]$ and $[\bullet,[\bullet]]$.

Definition 1.2 (Elementary Differentials). For a tree $\tau \in T$ the *elementary differential* is a mapping $F(\tau):\mathbb{R}^n \to \mathbb{R}^n$, defined recursively by $F(\bullet)(y) = f(y)$ and

$$F(\tau)(y) = f^{(m)}(y)\Big(F(\tau_1)(y),\ldots,F(\tau_m)(y)\Big) \qquad \text{for} \quad \tau = [\tau_1,\ldots,\tau_m].$$

Examples of these constructions and the corresponding coefficients are seen in Table 1.1. With these definitions, we obtain from (1.4):

Theorem 1.3. *The qth derivative of the exact solution is given by*

$$y^{(q)}(t_0) = \sum_{|\tau|=q} \alpha(\tau)\, F(\tau)(y_0), \tag{1.5}$$

where $\alpha(\tau)$ are positive integer coefficients. □

Third Step. We now turn to the numerical solution of the Runge–Kutta method (II.1.4), which, by putting $hk_i = g_i$, we write as

$$g_i = hf(u_i) \tag{1.6}$$

and

$$u_i = y_0 + \sum_j a_{ij}\, g_j, \qquad y_1 = y_0 + \sum_i b_i\, g_i, \tag{1.7}$$

where u_i, g_i and y_1 are functions of h. We develop the derivatives of (1.6), by Leibniz' rule, and obtain $g_i^{(q)} = h(f(u_i))^{(q)} + q \cdot (f(u_i))^{(q-1)}$. This gives, for $h = 0$,

$$g_i^{(q)} = q \cdot (f(u_i))^{(q-1)}, \tag{1.8}$$

the same expression as in (1.2), with y just replaced by u_i and with an extra factor q. Consequently, exactly as in (1.3),

$$\begin{aligned}
\dot g_i &= 1 \cdot f(y_0) \\
\ddot g_i &= 2 \cdot f'(y_0)\, \dot u_i \qquad\qquad (1.9)\\
g_i^{(3)} &= 3 \cdot \big(f''(y_0)(\dot u_i, \dot u_i) + f'(y_0)\, \ddot u_i \big) \\
g_i^{(4)} &= 4 \cdot \big(f'''(y_0)(\dot u_i, \dot u_i, \dot u_i) + 3f''(y_0)(\ddot u_i, \dot u_i) + f'(y_0)\, u_i^{(3)} \big) \\
g_i^{(5)} &= 5 \cdot \big(f^{(4)}(y_0)(\dot u_i, \dot u_i, \dot u_i, \dot u_i) + 6f'''(y_0)(\ddot u_i, \dot u_i, \dot u_i) + 4f''(y_0)(u_i^{(3)}, \dot u_i) \\
&\qquad + 3f''(y_0)(\ddot u_i, \ddot u_i) + f'(y_0)\, u_i^{(4)} \big),
\end{aligned}$$

and so on. Here, the derivatives of g_i and u_i are evaluated at $h = 0$.

Fourth Step. We now insert recursively the derivatives $\dot u_i$, $\ddot u_i, \ldots$ into (1.9). This will give the next higher derivative of g_i, and, using

$$u_i^{(q)} = \sum_j a_{ij} \cdot g_j^{(q)}, \tag{1.10}$$

which follows from (1.7), also the next higher derivative of u_i. This process begins as

$$\begin{aligned}
\dot g_i &= 1 \cdot f & \dot u_i &= 1 \cdot \big(\textstyle\sum_j a_{ij}\big) \cdot f \\
\ddot g_i &= (1 \cdot 2)\, \big(\textstyle\sum_j a_{ij}\big) f'f & \ddot u_i &= (1 \cdot 2)\, \big(\textstyle\sum_{jk} a_{ij} a_{jk}\big) f'f
\end{aligned} \tag{1.11}$$

and so on. If we compare these formulas with the first lines of (1.4), we see that the results are precisely the same, apart from the extra factors. We denote the *integer factors* $1, 1 \cdot 2, \ldots$ by $\gamma(\tau)$ and the factors containing the a_{ij}'s by $\mathbf{g}_i(\tau)$ and $\mathbf{u}_i(\tau)$, respectively. We obtain by induction that the same happens in general, i.e. that, in contrast to (1.5),

$$
\begin{aligned}
g_i^{(q)}\big|_{h=0} &= \sum_{|\tau|=q} \gamma(\tau)\cdot \mathbf{g}_i(\tau)\cdot \alpha(\tau)\, F(\tau)(y_0) \\
u_i^{(q)}\big|_{h=0} &= \sum_{|\tau|=q} \gamma(\tau)\cdot \mathbf{u}_i(\tau)\cdot \alpha(\tau)\, F(\tau)(y_0),
\end{aligned}
\tag{1.12}
$$

where $\alpha(\tau)$ and $F(\tau)$ are *the same* quantities as before. This is seen by continuing the insertion process of the derivatives $u_i^{(q)}$ into the right-hand side of (1.9). For example, if $\dot u_i$ and $\ddot u_i$ are inserted into $3f''(\ddot u_i, \dot u_i)$, we will obtain the corresponding expression as in (1.4), multiplied by the two extra factors $\mathbf{u}_i(\,[\bullet]\,)$, brought in by $\ddot u_i$, and $\mathbf{u}_i(\,\bullet\,)$ from $\dot u_i$. For a general tree $\tau = [\tau_1, \ldots, \tau_m]$ this will be

$$
\mathbf{g}_i(\tau) = \mathbf{u}_i(\tau_1)\cdot \ldots \cdot \mathbf{u}_i(\tau_m)\ . \tag{1.13}
$$

Second, the factors $\gamma(\,[\bullet]\,)$ and $\gamma(\,\bullet\,)$ will receive the additional factor $q = |\tau|$ from (1.9), i.e., we will have in general

$$
\gamma(\tau) = |\tau|\, \gamma(\tau_1)\cdot \ldots \cdot \gamma(\tau_m). \tag{1.14}
$$

Then, by (1.10),

$$
\mathbf{u}_i(\tau) = \sum_j a_{ij}\, \mathbf{g}_j(\tau) = \sum_j a_{ij}\cdot \mathbf{u}_j(\tau_1)\cdot \ldots \cdot \mathbf{u}_j(\tau_m). \tag{1.15}
$$

This formula can be re-used repeatedly, as long as some of the trees $\tau_1, \ldots, \tau_m$ are of order > 1. Finally, we have from the last formula of (1.7), that the coefficients for the numerical solution, which we denote by $\phi(\tau)$ and call the *elementary weights*, satisfy

$$
\phi(\tau) = \sum_i b_i\, \mathbf{g}_i(\tau). \tag{1.16}
$$

We summarize the result as follows:

Theorem 1.4. *The derivatives of the numerical solution of a Runge–Kutta method (II.1.4), for $h = 0$, are given by*

$$
y_1^{(q)}\big|_{h=0} = \sum_{|\tau|=q} \gamma(\tau)\cdot \phi(\tau)\cdot \alpha(\tau)\, F(\tau)(y_0), \tag{1.17}
$$

where $\alpha(\tau)$ and $F(\tau)$ are the same as in Theorem 1.3, the coefficients $\gamma(\tau)$ satisfy $\gamma(\,\bullet\,) = 1$ and (1.14). The elementary weights $\phi(\tau)$ are obtained from the tree τ as follows: attach to every vertex a summation letter ("i" to the root), then $\phi(\tau)$ is the sum, over all summation indices, of a product composed of b_i, and factors a_{jk} for each vertex "j" directly connected with "k" by an upwards directed branch.

Proof. Repeated application of (1.15) followed by (1.16) shows that the elementary weight $\phi(\tau)$ is the collection of $\sum_i b_i$ from (1.16) and all $\sum_j a_{ij}$ of (1.15). □

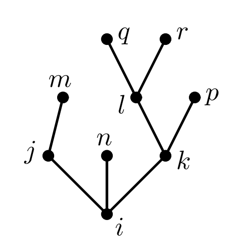
Theorem 1.5. *The Runge–Kutta method has order p if and only if*

$$\phi(\tau) = \frac{1}{\gamma(\tau)} \qquad \textit{for} \quad |\tau| \le p. \tag{1.18}$$

Proof. The comparison of Theorem 1.3 with Theorem 1.4 proves the sufficiency of condition (1.18). The necessity of (1.18) follows from the independence of the elementary differentials (see e.g., Hairer, Nørsett & Wanner (1993), Exercise 4 of Sect. II.2). □

Example 1.6. For the following tree of order 9 we have

$$\sum_{i,j,k,l,m,n,p,q,r} b_i a_{ij} a_{jm} a_{in} a_{ik} a_{kl} a_{lq} a_{lr} a_{kp} = \frac{1}{9 \cdot 2 \cdot 5 \cdot 3}$$



or, by using $\sum_j a_{ij} = c_i$,

$$\sum_{i,j,k,l} b_i c_i a_{ij} c_j a_{ik} c_k a_{kl} c_l^2 = \frac{1}{270}.$$

The quantities $\phi(\tau)$ and $\gamma(\tau)$ for all trees up to order 4 are given in Table 1.1. This also verifies the formulas (II.1.6) stated previously.

III.1.2 B-Series

We now introduce the concept of B-series, which gives further insight into the behaviour of numerical methods and allows extensions to more general classes of methods.

Motivated by formulas (1.12) and (1.17) above, we consider the corresponding *series* as the objects of our study. This means, we study power series in $h^{|\tau|}$ containing elementary differentials $F(\tau)$ and arbitrary coefficients which are now written in the form $a(\tau)$. Such series will be called B-series. To move from (1.6) to (1.13) we need to prove a result stating that *a B-series inserted into $hf(\cdot)$ is again a B-series*. We start with

$$B(a,y) = y + a(\bullet)hf(y) + a([\bullet])h^2(f'f)(y) + \ldots = y + \delta, \tag{1.19}$$

and get by Taylor expansion

$$hf\big(B(a,y)\big) = hf(y+\delta) = hf(y) + hf'(y)\delta + \frac{h}{2!}f''(y)(\delta,\delta) + \ldots . \tag{1.20}$$

Inserting δ from (1.19) and multiplying out, we obtain the expression

$$\begin{aligned} hf\big(B(a,y)\big) = {} & hf + h^2 a(\bullet) f'f + h^3 a([\bullet]) f'f'f + \frac{h^3}{2!} a(\bullet)^2 f''(f,f) \\ & + h^4 a(\bullet) a([\bullet]) f''(f'f,f) + \ldots . \end{aligned} \tag{1.21}$$

This beautiful formula is not yet perfect for two reasons. First, there is a denominator 2! in the fourth term. The origin of this lies in the *symmetry* of the tree $\vee$. We thus introduce the symmetry coefficients of Definition 1.7 (following Butcher 1987, Theorem 144A). Second, there is no first term y. We therefore allow the factor $a(\emptyset)$ in Definition 1.8.

Definition 1.7 (Symmetry coefficients). The symmetry coefficients $\sigma(\tau)$ are defined by $\sigma(\bullet) = 1$ and, for $\tau = [\tau_1, \ldots, \tau_m]$,

$$\sigma(\tau) = \sigma(\tau_1) \cdot \ldots \cdot \sigma(\tau_m) \cdot \mu_1! \mu_2! \cdot \ldots \ , \tag{1.22}$$

where the integers $\mu_1, \mu_2, \ldots$ count equal trees among $\tau_1, \ldots, \tau_m$.

Definition 1.8 (B-Series). For a mapping $a : T \cup \{\emptyset\} \to \mathbb{R}$ a formal series of the form

$$B(a, y) = a(\emptyset) y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \, a(\tau) \, F(\tau)(y) \tag{1.23}$$

is called a *B-series*.[1]

The main results of the theory of B-series have their origin in the paper of Butcher (1972), although series expansions were not used there. B-series were then introduced by Hairer & Wanner (1974). The normalization used in Definition 1.8 is due to Butcher & Sanz-Serna (1996). The following fundamental lemma gives a second way of finding the order conditions.

Lemma 1.9. *Let $a : T \cup \{\emptyset\} \to \mathbb{R}$ be a mapping satisfying $a(\emptyset) = 1$. Then the corresponding B-series inserted into $hf(\cdot)$ is again a B-series. That is*

$$h\, f\Big(B(a, y)\Big) = B(a', y), \tag{1.24}$$

where $a'(\emptyset) = 0$, $a'(\bullet) = 1$, and

$$a'(\tau) = a(\tau_1) \cdot \ldots \cdot a(\tau_m) \qquad \textit{for} \quad \tau = [\tau_1, \ldots, \tau_m]. \tag{1.25}$$

Proof. Since $a(\emptyset) = 1$ we have $B(a, y) = y + \mathcal{O}(h)$, so that $hf\big(B(a, y)\big)$ can be expanded into a Taylor series around y. As in formulas (1.20) and (1.21), we get

[1] In this section we are not concerned about the convergence of the series. We shall see later in Chap. IX that the series converges for sufficiently small h, if $a(\tau)$ satisfies an inequality $|a(\tau)| \le \gamma(\tau) c d^{|\tau|}$ and if $f(y)$ is an analytic function. If $f(y)$ is only k-times differentiable, then all formulas of this section remain valid for the truncated B-series $\sum_{\tau \in T, |\tau| \le k} \cdot / \cdot$ with a suitable remainder term of size $\mathcal{O}(h^{k+1})$ added.

$$
\begin{aligned}
hf\big(B(a,y)\big) &= h\sum_{m\geq 0}\frac{1}{m!}\,f^{(m)}(y)\Big(B(a,y)-y\Big)^m\\
&= h\sum_{m\geq 0}\frac{1}{m!}\sum_{\tau_1\in T}\cdots\sum_{\tau_m\in T}\frac{h^{|\tau_1|+\ldots+|\tau_m|}}{\sigma(\tau_1)\cdot\ldots\cdot\sigma(\tau_m)}\;\cdot a(\tau_1)\cdot\ldots\cdot a(\tau_m)\\
&\qquad\qquad\qquad\qquad\cdot f^{(m)}(y)\Big(F(\tau_1)(y),\ldots,F(\tau_m)(y)\Big)\\
&= \sum_{m\geq 0}\sum_{\tau_1\in T}\cdots\sum_{\tau_m\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\,\frac{\mu_1!\mu_2!\cdot\ldots}{m!}\cdot a'(\tau)\,F(\tau)(y)\\
&\qquad\qquad\qquad\qquad\text{with } \tau=[\tau_1,\ldots,\tau_m]\\
&= \sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\,a'(\tau)\,F(\tau)(y) \;=\; B(a',y).
\end{aligned}
$$

The last equality follows from the fact that there are $\binom{m}{\mu_1,\mu_2,\ldots}$ possibilities for writing the tree τ in the form $\tau=[\tau_1,\ldots,\tau_m]$. For example, the trees $[\bullet,\bullet,[\bullet]]$, $[\bullet,[\bullet],\bullet]$ and $[[\bullet],\bullet,\bullet]$ appear as different terms in the upper sum, but only as one term in the lower sum. □

Back to the Order Conditions. We present now a new derivation of the order conditions that is solely based on B-series and on Lemma 1.9. Let a Runge–Kutta method, say formulas (1.6) and (1.7), be given. All quantities in the defining formulas are set up as B-series, $g_i=B(\mathbf{g}_i,y_0)$, $u_i=B(\mathbf{u}_i,y_0)$, $y_1=B(\phi,y_0)$. Then, either the linearity and/or Lemma 1.9, translate the formulas of the method into corresponding formulas for the coefficients (1.13), (1.15), and (1.16). This recursively justifies the ansatz as B-series.

Assuming the *exact* solution to be a B-series $B(\mathbf{e},y_0)$, a term-by-term derivation of this series and an application of Lemma 1.9 to (1.1) yields

$$
\mathbf{e}(\tau)=\frac{1}{|\tau|}\,\mathbf{e}(\tau_1)\cdot\ldots\cdot\mathbf{e}(\tau_m).
$$

Together with definition (1.14) of $\gamma(\tau)$ we thus obtain

$$
\mathbf{e}(\tau)=\frac{1}{\gamma(\tau)}. \tag{1.26}
$$

A comparison of the coefficients of the B-series $y_1=B(\varphi,y_0)$ with those of the exact solution gives (1.18) and proves Theorem 1.5 again .

Comparing the B-series $B(\mathbf{e},y_0)$ for the exact solution with Theorem 1.3, we get as a byproduct the formula

$$
\alpha(\tau)=\frac{|\tau|!}{\sigma(\tau)\cdot\gamma(\tau)}. \tag{1.27}
$$

If the available tools are enriched by the more general composition law of Theorem 1.10 below, this procedure can be applied to yet larger classes of methods.

III.1.3 Composition of Methods

The order theory for the composition of methods goes back to 1969, when Butcher used it to circumvent the order barrier for explicit 5th order 5 stage methods. It led to the seminal publication of Butcher (1972), where the general composition formula in (1.34) was expressed recursively.

Composition of Runge–Kutta Methods. Suppose that, starting from an initial value y_0, we compute a numerical solution y_1 using a Runge–Kutta method with coefficients a_{ij}, b_i and step size h. Then, continuing from y_1, we compute a value y_2 using another method with coefficients a^*_{ij}, b^*_i and the same step size. This composition of two methods is now considered as a *single* method (with coefficients $\widehat{a}_{ij}, \widehat{b}_i$). The problem is to derive the order properties of this new method, in particular to express the elementary weights $\widehat{\phi}(\tau)$ in terms of those of the original two methods.

If the value y_1 from the first method is inserted into the starting value for the second method, one sees that the coefficients of the combined method are given by (here written for two-stage methods)

$$\begin{array}{|cccc} \widehat{a}_{11} & \widehat{a}_{12} & & \\ \widehat{a}_{21} & \widehat{a}_{22} & & \\ \widehat{a}_{31} & \widehat{a}_{32} & \widehat{a}_{33} & \widehat{a}_{34} \\ \widehat{a}_{41} & \widehat{a}_{42} & \widehat{a}_{43} & \widehat{a}_{44} \\ \hline \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widehat{b}_4 \end{array} \quad = \quad \begin{array}{|cccc} a_{11} & a_{12} & & \\ a_{21} & a_{22} & & \\ b_1 & b_2 & a^*_{11} & a^*_{12} \\ b_1 & b_2 & a^*_{21} & a^*_{22} \\ \hline b_1 & b_2 & b^*_1 & b^*_2 \end{array} \tag{1.28}$$

and our problem is to compute the elementary weights of this scheme.

Derivation. The idea is to write the sum for $\widehat{\phi}(\tau)$, say for the tree $[[\bullet],\bullet]$, in full detail

$$\widehat{\phi}([[\bullet],\bullet]) = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{k=1}^{4}\sum_{l=1}^{4} \widehat{b}_i\, \widehat{a}_{ij}\, \widehat{a}_{ik}\, \widehat{a}_{kl} = \ldots \tag{1.29}$$

and to split each sum into the two different index sets. This leads to $2^{|\tau|}$ different expressions $\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \cdot/\cdot + \sum_{i=3}^{4}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \cdot/\cdot + \sum_{i=1}^{2}\sum_{j=3}^{4}\sum_{k=1}^{2}\sum_{l=1}^{2} \cdot/\cdot + \ldots$. We symbolize each expression by drawing the corresponding vertex of τ as a *bullet* for the first index set and as a *star* for the second. However, due to the zero pattern in the matrix in (1.28) (the upper right corner is missing), each term with "star above bullet" can be omitted, since the corresponding $\widehat{a}_{ij}$'s are zero. So the only combinations to be considered are those of Fig. 1.1. We finally insert the quantities from the right tableau in (1.28),

$$\begin{aligned} \widehat{\phi}([[\bullet],\bullet]) &= \sum b_i\, a_{ij}\, a_{ik}\, a_{kl} + \sum b^*_i\, b_j\, b_k\, a_{kl} + \sum b^*_i\, a^*_{ij}\, b_k\, a_{kl} + \sum b^*_i\, b_j\, a^*_{ik}\, b_l \\ &\quad + \sum b^*_i\, a^*_{ij}\, a^*_{ik}\, b_l + \sum b^*_i\, b_j\, a^*_{ik}\, a^*_{kl} + \sum b^*_i\, a^*_{ij}\, a^*_{ik}\, a^*_{kl}, \end{aligned}$$

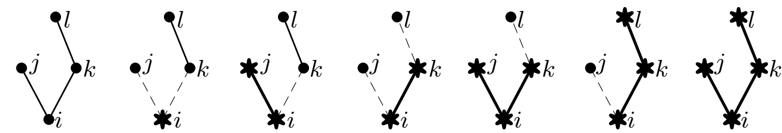


Fig. 1.1. Combinations with nonzero product

and we observe that each factor of the type b_j interrupts the summation, so that the terms decompose into factors of elementary weights of the individual methods as follows:

$$\begin{aligned}\widehat{\phi}([\bullet,[\bullet]]) &= \phi([\bullet,[\bullet]]) + \phi^*(\bullet)\cdot\phi(\bullet)\phi([\bullet]) + \phi^*([\bullet])\cdot\phi([\bullet]) + \phi^*([\bullet])\cdot\phi(\bullet)\phi(\bullet)\\ &\quad + \phi^*([\bullet,\bullet])\cdot\phi(\bullet) + \phi^*([[\bullet]])\cdot\phi(\bullet) + \phi^*([\bullet,[\bullet]])\,.\end{aligned}$$

The trees composed of the "star" nodes of τ in Fig. 1.1 constitute all possible "subtrees" θ (from the empty tree to τ itself) having the same root as τ. This is the key for understanding the general result.

Ordered Trees. In order to formalize the procedure of Fig. 1.1, we introduce the set *OT* of *ordered trees* recursively as follows: $\bullet\in$ *OT*, and

$$\text{if } \omega_1,\ldots,\omega_m \in \mathit{OT}, \text{ then also the ordered } m\text{-tuple } (\omega_1,\ldots,\omega_m)\in \mathit{OT}. \tag{1.30}$$

As the name suggests, in the graphical representation of an ordered tree the order of the branches leaving cannot be permuted. Neglecting the ordering, a tree $\tau\in T$ can be considered as an equivalence class of ordered trees, denoted $\tau=\overline{\omega}$.

For example, the tree of Fig. 1.1 has two orderings, namely $(\bullet,(\bullet))$ and $((\bullet),\bullet)$. We denote by $\nu(\tau)$ the number of possible orderings of the tree τ. It is given by $\nu(\bullet)=1$ and

$$\nu(\tau)=\frac{m!}{\mu_1!\mu_2!\cdot\ldots}\,\nu(\tau_1)\cdot\ldots\cdot\nu(\tau_m) \tag{1.31}$$

for $\tau=[\tau_1,\ldots,\tau_m]$, where the integers $\mu_1,\mu_2,\ldots$ are the numbers of equal trees among $\tau_1,\ldots,\tau_m$. This number is closely related to the symmetry coefficient $\sigma(\tau)$, because the product $\kappa(\tau)=\sigma(\tau)\nu(\tau)$ satisfies the recurrence relation

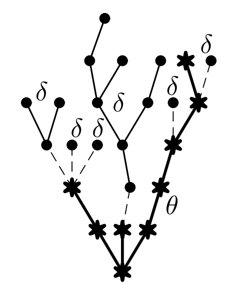
$$\kappa(\tau)=m!\,\kappa(\tau_1)\cdot\ldots\cdot\kappa(\tau_m). \tag{1.32}$$

We introduce the set *OST*(ω) of *ordered subtrees* of an ordered tree $\omega\in$ *OT* by

$$\begin{aligned}\mathit{OST}(\bullet) &= \{\emptyset,\bullet\} \\ \mathit{OST}(\omega) &= \{\emptyset\}\cup\{(\theta_1,\ldots,\theta_m)\;;\;\theta_i\in \mathit{OST}(\omega_i)\} \quad \text{for } \omega=(\omega_1,\ldots,\omega_m).\end{aligned}\tag{1.33}$$

Each ordered subtree $\theta\in$ *OST*(ω) is naturally associated with a tree $\overline{\theta}\in T$ obtained by neglecting the ordering and the $\emptyset$-components of θ. For every tree $\tau\in T$ we choose, once and for all, an ordering. We denote this ordered tree by $\omega(\tau)$, and we put *OST*(τ) = *OST*$(\omega(\tau))$.

For the tree of Fig. 1.1, considered as an ordered tree, the ordered subtrees correspond to the trees composed of the "star" nodes.



The General Rule. The general composition rule now becomes visible: for $\theta \in OST(\omega)$ we denote by $\omega \setminus \theta$ the "forest" collecting the trees left over when θ has been removed from the ordered tree ω. For brevity we set $\tau\setminus\theta := \omega(\tau)\setminus\theta$. With the conventions $\phi^*(\theta) = \phi^*(\overline{\theta})$ and $\phi^*(\emptyset) = 1$ we then have

$$\widehat{\phi}(\tau) = \sum_{\theta \in OST(\tau)} \Big(\phi^*(\theta) \cdot \prod_{\delta \in \tau\setminus\theta} \phi(\delta) \Big) . \qquad (1.34)$$

This composition formula for the trees up to order 3 reads:

$$\begin{aligned}
\widehat{\phi}(\bullet) &= \phi^*(\emptyset)\cdot\phi(\bullet) + \phi^*(\bullet)\\
\widehat{\phi}([\bullet]) &= \phi^*(\emptyset)\cdot\phi([\bullet]) + \phi^*(\bullet)\cdot\phi(\bullet) + \phi^*([\bullet])\\
\widehat{\phi}([\bullet,\bullet]) &= \phi^*(\emptyset)\cdot\phi([\bullet,\bullet]) + \phi^*(\bullet)\cdot\phi(\bullet)^2 + 2\phi^*([\bullet])\cdot\phi(\bullet) + \phi^*([\bullet,\bullet])\\
\widehat{\phi}([[\bullet]]) &= \phi^*(\emptyset)\cdot\phi([[\bullet]]) + \phi^*(\bullet)\cdot\phi([\bullet]) + \phi^*([\bullet])\cdot\phi(\bullet) + \phi^*([[\bullet]])
\end{aligned}$$

The tree $\tau = [\bullet,\bullet]$ has the subtrees displayed in Fig. 1.2. It contains symmetries in that the third and fourth subtrees are topologically equivalent. This explains the factor 2 in the expression for the elementary weight.

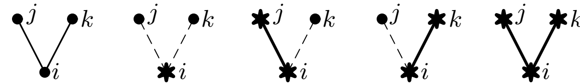


Fig. 1.2. A tree with symmetry

III.1.4 Composition of B-Series

We now extend the above composition law to general B-series, i.e., we insert the B-series themselves into each other, as sketched in Fig. 1.3. This allows us to generalize Lemma 1.9 (because $hf(y)$ is a special B-series).

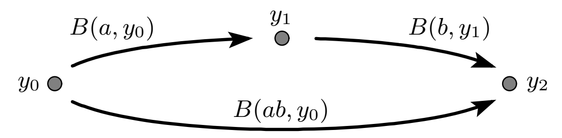


Fig. 1.3. Composition of B-series

We start with an observation of Murua (see, e.g., Murua & Sanz-Serna (1999), p. 1083), namely that the proof of Lemma 1.9 remains the same if the function $hf(y)$ is replaced with any other function $hg(y)$; in this case (1.21) is replaced with

$$
\begin{aligned}
hg\big(B(a,y)\big) = hg + h^2 a(\bullet) g'f + h^3 a\big(\substack{\bullet\\ |\\ \bullet}\big) g'f'f + \frac{h^3}{2!} a(\bullet)^2 g''(f,f) \\
+ h^4 a(\bullet) a\big(\substack{\bullet\\ |\\ \bullet}\big) g''(f'f,f) + \ldots \, .
\end{aligned}
\tag{1.35}
$$

Such series will reappear in Sect. III.3.1 below. Extending this idea further to, say, $f''(y)(v_1,v_2)$, where v_1, v_2 are two fixed vectors, we obtain

$$
\begin{aligned}
hf''\big(B(a,y)\big)(v_1,v_2) &= hf''(v_1,v_2) + h^2 a(\bullet) f'''(v_1,v_2,f) \\
&+ h^3 a\big(\substack{\bullet\\ |\\ \bullet}\big) f'''(v_1,v_2,f'f) + \frac{1}{2!} h^3 a(\bullet)^2 f''''(v_1,v_2,f,f) \\
&+ h^4 a(\bullet) a\big(\substack{\bullet\\ |\\ \bullet}\big) f''''(v_1,v_2,f'f,f) + \ldots \, .
\end{aligned}
\tag{1.36}
$$

This idea will lead to a direct proof of the following theorem of Hairer & Wanner (1974).

Theorem 1.10. *Let* $a : T \cup \{\emptyset\} \to \mathbb{R}$ *be a mapping satisfying* $a(\emptyset) = 1$ *and let* $b : T \cup \{\emptyset\} \to \mathbb{R}$ *be arbitrary. Then the B-series* $B(a,y)$ *inserted into* $B(b,\cdot)$ *is again a B-series*

$$
B\Big(b, B(a,y)\Big) = B(ab,y), \tag{1.37}
$$

where the group operation $ab(\tau)$ *is as in (1.34), i.e.,*

$$
ab(\tau) = \sum_{\theta \in OST(\tau)} b(\theta) \cdot a(\tau \setminus \theta) \qquad \textit{with} \qquad a(\tau \setminus \theta) = \prod_{\delta \in \tau \setminus \theta} a(\delta). \tag{1.38}
$$

Proof. (a) In part (c) below we prove by induction on $|\vartheta|$, $\vartheta \in T$ that

$$
\frac{h^{|\vartheta|}}{\sigma(\vartheta)} F(\vartheta)\Big(B(a,y)\Big) = \sum_{(\tau,\theta)\in A(\vartheta)} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau\setminus\theta)\, F(\tau)(y), \tag{1.39}
$$

where

$$
A(\vartheta) = \big\{(\tau,\theta)\,;\ \tau\in T, \theta \in OST(\tau), \overline{\theta} = \vartheta\big\}.
$$

Multiplying (1.39) by $b(\vartheta)$ and summing over all $\vartheta \in T$ yields the statement (1.37)-(1.38), because

$$
\sum_{\vartheta\in T}\ \sum_{(\tau,\theta)\in A(\vartheta)} \cdot/\cdot = \sum_{\tau\in T}\ \sum_{\theta\in OST(\tau)} \cdot/\cdot\, .
$$

(b) Choosing a different ordering of τ in the definition of $OST(\tau)$ yields the same sum in (1.39). Therefore (1.39) is equivalent to

$$
\frac{h^{|\vartheta|}}{\sigma(\vartheta)} F(\vartheta)\Big(B(a,y)\Big) = \sum_{(\omega,\theta)\in \Omega(\vartheta)} \frac{h^{|\omega|}}{\sigma(\omega)\nu(\omega)} a(\omega\setminus\theta)\, F(\omega)(y), \tag{1.40}
$$

where

$$
\Omega(\vartheta) = \big\{(\omega,\theta)\,;\ \omega\in OT, \theta \in OST(\omega), \overline{\theta} = \vartheta\big\},
$$

and $\nu(\tau)$ is the number of orderings of the tree τ, see (1.31). Functions defined on trees are naturally extended to ordered trees. In (1.40) we use $|\omega| = |\tau|$, $\sigma(\omega) = \sigma(\tau)$, $\nu(\omega) = \nu(\tau)$, $a(\omega\setminus\theta) = a(\tau\setminus\theta)$, and $F(\omega)(y) = F(\tau)(y)$ for $\overline{\omega} = \tau$.

(c) For $\vartheta = \bullet$ and $\omega = (\omega_1, \ldots, \omega_m)$ we have $a(\omega \setminus \theta) = a(\omega_1) \cdot \ldots \cdot a(\omega_m)$ if $\overline{\theta} = \bullet$. Since we have a one-to-one correspondence $(\omega, \theta) \leftrightarrow \omega$ between $\Omega(\bullet)$ and *OT*, and since the expression in the sum of (1.40) is independent of the ordering of ω, formula (1.40) is precisely Lemma 1.9.

To prove (1.40) for a general tree $\vartheta = [\vartheta_1, \ldots, \vartheta_l]$, we apply the idea put forward in (1.36) to $hf^{(l)}\big(B(a,y)\big)(v_1, \ldots, v_l)$ with fixed $v_1, \ldots, v_l$, and obtain as in the proof of Lemma 1.9

$$hf^{(l)}\big(B(a,y)\big)(v_1,\ldots,v_l) = \sum_{m\ge 0} \frac{1}{m!} \sum_{\tau_{l+1}\in T} \cdots \sum_{\tau_{l+m}\in T} \frac{h^{|\tau_{l+1}|+\ldots+|\tau_{l+m}|+1}}{\sigma(\tau_{l+1})\cdot\ldots\cdot\sigma(\tau_{l+m})}$$
$$\cdot\, a(\tau_{l+1})\cdot\ldots\cdot a(\tau_{l+m})\cdot f^{(l+m)}(y)\Big(v_1,\ldots,v_l,F(\tau_{l+1})(y),\ldots,F(\tau_{l+m})(y)\Big).$$

Changing the sums over trees to sums over ordered trees we obtain

$$hf^{(l)}\big(B(a,y)\big)(v_1,\ldots,v_l) = \sum_{m\ge 0} \frac{1}{m!} \sum_{\omega_{l+1}\in OT} \cdots \sum_{\omega_{l+m}\in OT} \frac{h^{|\omega_{l+1}|+\ldots+|\omega_{l+m}|+1}}{\kappa(\omega_{l+1})\cdot\ldots\cdot\kappa(\omega_{l+m})}$$
$$\cdot\, a(\omega_{l+1})\cdot\ldots\cdot a(\omega_{l+m})\cdot f^{(l+m)}(y)\Big(v_1,\ldots,v_l,F(\omega_{l+1})(y),\ldots,F(\omega_{l+m})(y)\Big).$$

We insert $v_j = \frac{h^{|\vartheta_j|}}{\sigma(\vartheta_j)}\, F(\vartheta_j)\big(B(a,y)\big)$ into this relation, and we apply our induction hypothesis

$$v_j = \frac{h^{|\vartheta_j|}}{\sigma(\vartheta_j)}\, F(\vartheta_j)\Big(B(a,y)\Big) = \sum_{(\omega_j,\theta_j)\in\Omega(\vartheta_j)} \frac{h^{|\omega_j|}}{\kappa(\omega_j)}\, a(\omega_j\setminus\theta_j)\, F(\omega_j)(y).$$

We then use the recursive definitions of $\sigma(\vartheta)$ and $F(\vartheta)(y)$ on the left-hand side. On the right-hand side we use the multilinearity of $f^{(l+m)}$, the recursive definitions of $|\omega|$, $\kappa(\omega)$, $F(\omega)(y)$ for $\omega = (\omega_1, \ldots, \omega_{l+m})$, and the facts that

$$a(\omega\setminus\theta) = a(\omega_1\setminus\theta_1)\cdot\ldots\cdot a(\omega_l\setminus\theta_l)\cdot a(\omega_{l+1})\cdot\ldots\cdot a(\omega_{l+m})$$

and

$$\sum_{(\omega_1,\theta_1)\in\Omega(\vartheta_1)} \cdots \sum_{(\omega_l,\theta_l)\in\Omega(\vartheta_l)} \sum_{\omega_{l+1}\in OT} \cdots \sum_{\omega_{l+m}\in OT} \cdot/\cdot = \frac{m!\mu_1!\mu_2!\cdot\ldots}{(l+m)!} \sum_{(\omega,\theta)\in\Omega_{l+m}(\vartheta)} \cdot/\cdot$$

where $\mu_1, \mu_2, \ldots$ count equal trees among $\vartheta_1, \ldots, \vartheta_l$, and $\Omega_{l+m}(\vartheta)$ consists of those pairs $(\omega,\theta) \in \Omega(\vartheta)$ for which ω is of the form $\omega = (\omega_1, \ldots, \omega_{l+m})$. The factorials appear, because to every $(l+m)$-tuple of the left-hand sum correspond $\binom{l+m}{m,\mu_1,\mu_2,\ldots}$ elements in $\Omega_{l+m}(\vartheta)$, obtained by permuting the order. This yields formula (1.40) and hence (1.39). $\square$

Example 1.11. The composition laws for the trees of order ≤ 4 are

$$
\begin{aligned}
ab(\bullet) &= b(\emptyset)\cdot a(\bullet) + b(\bullet)\\
ab([\bullet]) &= b(\emptyset)\cdot a([\bullet]) + b(\bullet)\cdot a(\bullet) + b([\bullet])\\
ab([\bullet,\bullet]) &= b(\emptyset)\cdot a([\bullet,\bullet]) + b(\bullet)\cdot a(\bullet)^2 + 2b([\bullet])\cdot a(\bullet) + b([\bullet,\bullet])\\
ab([[\bullet]]) &= b(\emptyset)\cdot a([[\bullet]]) + b(\bullet)\cdot a([\bullet]) + b([\bullet])\cdot a(\bullet) + b([[\bullet]])\\
ab([\bullet,\bullet,\bullet]) &= b(\emptyset)\cdot a([\bullet,\bullet,\bullet]) + b(\bullet)\cdot a(\bullet)^3 + 3b([\bullet])\cdot a(\bullet)^2 + 3b([\bullet,\bullet])\cdot a(\bullet)\\
&\quad + b([\bullet,\bullet,\bullet])\\
ab([[\bullet],\bullet]) &= b(\emptyset)\cdot a([[\bullet],\bullet]) + b(\bullet)\cdot a(\bullet)a([\bullet]) + b([\bullet])\cdot a([\bullet]) + b([\bullet])\cdot a(\bullet)^2\\
&\quad + b([\bullet,\bullet])\cdot a(\bullet) + b([[\bullet]])\cdot a(\bullet) + b([[\bullet],\bullet])\\
ab([[\bullet,\bullet]]) &= b(\emptyset)\cdot a([[\bullet,\bullet]]) + b(\bullet)\cdot a([\bullet,\bullet]) + b([\bullet])\cdot a(\bullet)^2 + 2b([[\bullet]])\cdot a(\bullet)\\
&\quad + b([[\bullet,\bullet]])\\
ab([[[\bullet]]]) &= b(\emptyset)\cdot a([[[\bullet]]]) + b(\bullet)\cdot a([[\bullet]]) + b([\bullet])\cdot a([\bullet]) + b([[\bullet]])\cdot a(\bullet) + b([[[\bullet]]])
\end{aligned}
$$

Remark 1.12. The composition law (1.38) can alternatively be obtained from the corresponding formula (1.34) for Runge–Kutta methods by using the fact that B-series which represent Runge–Kutta methods are "dense" in the space of all B-series (see Theorem 306A of Butcher 1987).

III.1.5 The Butcher Group



John C. Butcher,
born: 31 March 1933 in Auckland
(New Zealand)

The composition law (1.38) can be turned into a *group operation*, by introducing a *unit element*

$$
e(\emptyset) = 1, \quad e(\tau) = 0 \quad \text{for } \tau \in T, \tag{1.41}
$$

and by computing the *inverse element* of a given a. This is obtained recursively from the table of Example 1.11, by requiring $aa^{-1}(\tau) = 0$ and by inserting the previously known values of $a^{-1}(\vartheta)$. This gives for the first orders

$$
\begin{aligned}
a^{-1}(\bullet) &= -a(\bullet)\\
a^{-1}([\bullet]) &= -a([\bullet]) + a(\bullet)^2\\
a^{-1}([\bullet,\bullet]) &= -a([\bullet,\bullet]) + 2a([\bullet])a(\bullet) - a(\bullet)^3\\
a^{-1}([[\bullet]]) &= -a([[\bullet]]) + 2a([\bullet])a(\bullet) - a(\bullet)^3
\end{aligned} \tag{1.42}
$$

We can distinguish several realizations of this group:

G_{RK} the set of Runge–Kutta schemes with composition (1.28);
G_{EW} the set of elementary weights of Runge–Kutta schemes with the composition law (1.34);
G_{TM} the set of tree mappings $a : T \cup \{\emptyset\} \to \mathbb{R}$ satisfying $a(\emptyset) = 1$ with composition (1.38);
G_{BS} the set of B-series (1.23) satisfying $a(\emptyset) = 1$ with composition (1.37).

A technical difficulty concerns the group G_{RK}, where "reducible" schemes must be identified (by deleting unnecessary stages or by combining stages that give identical results) to the same "irreducible" method (see Butcher (1972), or Butcher & Wanner (1996), p. 140). The definition of $\phi(\tau)$ in Theorem 1.4 describes a group isomorphism from G_{RK} to G_{EW}, further, G_{EW} is a subgroup of G_{TM} and Theorem 1.10 shows that formula (1.23) constitutes a group homomorphism from G_{TM} to G_{BS}. Because the elementary differentials are independent (see, e.g., Hairer, Nørsett & Wanner (1993), Exercise 4 of Sect. II.2), the last two groups are isomorphic. The group G_{RK} can also be extended by allowing "continuous" Runge–Kutta schemes with "infinitely many stages" (see Butcher (1972), or Butcher & Wanner (1996), p. 141). The term "Butcher group" was introduced by Hairer & Wanner (1974).

> This paper tells the story of a mathematical object that was created by John Butcher in 1972 and was rediscovered by Alain Connes, Henri Moscovici and Dirk Kreimer in 1998. (Ch. Brouder 2004)

Connection with Hopf Algebras and Quantum Field Theory. A surprising connection between Runge–Kutta theory and renormalization in quantum field theory has been discovered by Brouder (2000). One denotes by a *Hopf algebra* a graded algebra which, besides the usual product, also possesses a *coproduct*, a tool used by H. Hopf (1941) [2] in his topological classification of certain manifolds. Hopf algebras generated by families of rooted trees proved to be extremely useful for simplifying the intricate combinatorics of renormalization (Kreimer 1998). Kreimer's Hopf algebra $\mathcal{H}$ is the space generated by linear combinations of families of rooted trees and the coproduct is a mapping $\triangle : \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$ which is, for the first trees, given by

$$\begin{aligned} \triangle(\bullet) &= \bullet \otimes 1 + 1 \otimes \bullet \\ \triangle([\bullet]) &= [\bullet] \otimes 1 + \bullet \otimes \bullet + 1 \otimes [\bullet] \\ \triangle([\bullet,\bullet]) &= [\bullet,\bullet] \otimes 1 + \bullet\,\bullet \otimes \bullet + 2\,\bullet \otimes [\bullet] + 1 \otimes [\bullet,\bullet] \\ \triangle([[\bullet]]) &= [[\bullet]] \otimes 1 + [\bullet] \otimes \bullet + \bullet \otimes [\bullet] + 1 \otimes [[\bullet]] \end{aligned} \tag{1.43}$$

It can be clearly seen, that this algebraic structure is precisely the one underlying the composition law of Example 1.11, so that the Butcher group G_{TM} becomes the corresponding *character group*. The so-called *antipodes* of trees $\tau \in \mathcal{H}$, denoted by $S(\tau)$, are for the first trees

[2] Not to be confused with E. Hopf, the discoverer of the "Hopf bifurcation".

$$
\begin{aligned}
S(\bullet) &= -\bullet \\
S([\bullet]) &= -[\bullet] + \bullet\,\bullet \\
S([\bullet,\bullet]) &= -[\bullet,\bullet] + 2\,[\bullet]\,\bullet - \bullet\,\bullet\,\bullet \\
S([[\bullet]]) &= -[[\bullet]] + 2\,[\bullet]\,\bullet - \bullet\,\bullet\,\bullet
\end{aligned} \tag{1.44}
$$

and, apparently, describes the *inverse element* (1.42) in the Butcher group.

III.2 Order Conditions for Partitioned Runge–Kutta Methods

We now apply the ideas of the previous section to the creation of the order conditions for partitioned Runge–Kutta methods (II.2.2) of Sect. II.2. These results can then also be applied to Nyström methods.

III.2.1 Bi-Coloured Trees and P-Series

Let us consider a partitioned system

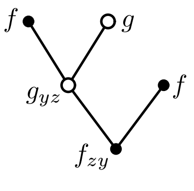
$$
\dot{y} = f(y,z), \qquad \dot{z} = g(y,z) \tag{2.1}
$$

(non-autonomous problems can be brought into this form by appending $\dot{t} = 1$). We start by computing the derivatives of its exact solution, which are to be inserted into the Taylor series expansion. By analogy with (1.4) we obtain in this case the derivatives of y at t_0 as follows:

$$
\begin{aligned}
\dot{y} &= f \\
\ddot{y} &= f_y f + f_z g \qquad\qquad (2.2) \\
y^{(3)} &= f_{yy}(f,f) + 2\, f_{yz}(f,g) + f_{zz}(g,g) + f_y f_y f + f_y f_z g + f_z g_y f + f_z g_z g.
\end{aligned}
$$

Here, $f_y, f_z, f_{yz}, \ldots$ denote partial derivatives and all terms are to be evaluated at (y_0, z_0). Similar expressions are obtained for the derivatives of $z(t)$.

The terms occurring in these expressions are again called the *elementary differentials* $F(\tau)(y,z)$. For their graphical representation as a tree τ, we distinguish between "black" vertices for representing an f and "white" vertices for a g. Upwards pointing branches represent partial derivatives, with respect to y if the branch leads to a black vertex, and with respect to z if it leads to a white vertex. With this convention, the graph to the right corresponds to the expression $f_{zy}\big(g_{yz}(f,g), f\big)$ (see Table 2.1 for more examples).



We denote by TP the set of graphs obtained by the above procedure, and we call them (rooted) *bi-coloured trees*. The first graphs are $\bullet$ and $\circ$. By analogy with Definition 1.1, we denote by

Table 2.1. Bi-coloured trees, elementary differentials, and coefficients

$\|\tau\|$	τ	graph	$\alpha(\tau)$	$F(\tau)$	$\gamma(\tau)$	$\phi(\tau)$	$\sigma(\tau)$
1	$\bullet$	$\bullet$	1	f	1	$\sum_i b_i$	1
2	$[\bullet]_y$		1	$f_y f$	2	$\sum_{ij} b_i a_{ij}$	1
2	$[\circ]_y$		1	$f_z g$	2	$\sum_{ij} b_i \widehat{a}_{ij}$	1
3	$[\bullet,\bullet]_y$		1	$f_{yy}(f,f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$	2
3	$[\bullet,\circ]_y$		2	$f_{yz}(f,g)$	3	$\sum_{ijk} b_i a_{ij} \widehat{a}_{ik}$	1
3	$[\circ,\circ]_y$		1	$f_{zz}(g,g)$	3	$\sum_{ijk} b_i \widehat{a}_{ij} \widehat{a}_{ik}$	2
3	$[[\bullet]_y]_y$		1	$f_y f_y f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$	1
3	$[[\circ]_y]_y$		1	$f_y f_z g$	6	$\sum_{ijk} b_i a_{ij} \widehat{a}_{jk}$	1
3	$[[\bullet]_z]_y$		1	$f_z g_y f$	6	$\sum_{ijk} b_i \widehat{a}_{ij} a_{jk}$	1
3	$[[\circ]_z]_y$		1	$f_z g_z g$	6	$\sum_{ijk} b_i \widehat{a}_{ij} \widehat{a}_{jk}$	1
1	$\circ$	$\circ$	1	g	1	$\sum_i \widehat{b}_i$	1
2	$[\bullet]_z$		1	$g_y f$	2	$\sum_{ij} \widehat{b}_i a_{ij}$	1
	etc	etc		etc		etc	

$$[\tau_1,\ldots,\tau_m]_y \qquad \text{and} \qquad [\tau_1,\ldots,\tau_m]_z, \qquad \tau_1,\ldots,\tau_m \in TP$$

the bi-coloured trees obtained by connecting the roots of $\tau_1,\ldots,\tau_m$ to a new root, which is $\bullet$ in the first case, and $\circ$ in the second. Furthermore, we denote by TP_y and TP_z the subsets of TP which are formed by trees with black and white roots, respectively. Hence, the trees of TP_y correspond to derivatives of $y(t)$, whereas those of TP_z correspond to derivatives of $z(t)$.

As in Definition 1.2 we denote the number of vertices of $\tau \in TP$ by $|\tau|$, the *order* of τ. The symmetry coefficient $\sigma(\tau)$ is again defined by

$$\sigma(\bullet) = \sigma(\circ) = 1,$$

and, for $\tau = [\tau_1,\ldots,\tau_m]_y$ or $\tau = [\tau_1,\ldots,\tau_m]_z$, by

$$\sigma(\tau) = \sigma(\tau_1)\cdot\ldots\cdot\sigma(\tau_m)\,\cdot\mu_1!\mu_2!\ldots, \tag{2.3}$$

where the integers $\mu_1,\mu_2,\ldots$ count equal trees among $\tau_1,\ldots,\tau_m \in TP$. This is formally the same definition as in Sect. III.1. Observe, however, that $\sigma(\tau)$ depends on the colouring of the vertices. For example, we have $\sigma(\ ) = 2$, but $\sigma(\ ) = 1$. By analogy with Definition 1.8 we have:

Definition 2.1 (P-Series). For a mapping $a : TP \cup \{\emptyset_y, \emptyset_z\} \to \mathbb{R}$ a series of the form

$$P\Big(a,(y,z)\Big) = \begin{pmatrix} a(\emptyset_y)y + \displaystyle\sum_{\tau \in TP_y} \frac{h^{|\tau|}}{\sigma(\tau)}\, a(\tau)\, F(\tau)(y,z) \\ a(\emptyset_z)z + \displaystyle\sum_{\tau \in TP_z} \frac{h^{|\tau|}}{\sigma(\tau)}\, a(\tau)\, F(\tau)(y,z) \end{pmatrix}$$

is called a *P-series*.

The following results correspond to Lemma 1.9 and formula (1.26). They are obtained in exactly the same manner as the corresponding results for non-partitioned Runge–Kutta methods (Sect. III.1). We therefore omit their proofs.

Lemma 2.2. *Let* $a : TP \cup \{\emptyset_y, \emptyset_z\} \to \mathbb{R}$ *satisfy* $a(\emptyset_y) = a(\emptyset_z) = 1$. *Then*

$$h \begin{pmatrix} f\big(P\big(a,(y,z)\big)\big) \\ g\big(P\big(a,(y,z)\big)\big) \end{pmatrix} = P\Big(a',(y,z)\Big),$$

where $a'(\emptyset_y) = a'(\emptyset_z) = 0$, $a'(\bullet) = a'(\circ) = 1$, *and*

$$a'(\tau) = a(\tau_1) \cdot \ldots \cdot a(\tau_m), \tag{2.4}$$

if either $\tau = [\tau_1, \ldots, \tau_m]_y$ *or* $\tau = [\tau_1, \ldots, \tau_m]_z$. □

Theorem 2.3 (P-Series of Exact Solution). *The exact solution of (2.1) is a P-series* $\big(y(t_0+h), z(t_0+h)\big) = P\big(\mathbf{e}, (y_0, z_0)\big)$, *where* $\mathbf{e}(\emptyset_y) = \mathbf{e}(\emptyset_z) = 1$ *and*

$$\mathbf{e}(\tau) = \frac{1}{\gamma(\tau)} \qquad \textit{for all} \quad t \in TP \tag{2.5}$$

where the $\gamma(\tau)$ *have the same values as for mono-coloured trees.* □

III.2.2 Order Conditions for Partitioned Runge–Kutta Methods

The next result corresponds to Theorem 1.4 and is a consequence of Lemma 2.2.

Theorem 2.4 (P-Series of Numerical Solution). *The numerical solution of a partitioned Runge–Kutta method (II.2.2) is a P-series* $(y_1, z_1) = P\big(\phi, (y_0, z_0)\big)$, *where* $\phi(\emptyset_y) = \phi(\emptyset_z) = 1$ *and*

$$\phi(\tau) = \begin{cases} \sum_{i=1}^s b_i \phi_i(\tau) & \textit{for} \quad \tau \in TP_y \\ \sum_{i=1}^s \widehat{b}_i \phi_i(\tau) & \textit{for} \quad \tau \in TP_z. \end{cases} \tag{2.6}$$

The expression $\phi_i(\tau)$ *is defined by* $\phi_i(\bullet) = \phi_i(\circ) = 1$ *and by*

$$\phi_i(\tau) = \psi_i(\tau_1) \cdot \ldots \cdot \psi_i(\tau_m) \quad \textit{with} \quad \psi_i(\tau_k) = \begin{cases} \sum_{j_k=1}^s a_{ij_k} \phi_{j_k}(\tau_k) & \textit{if } \tau_k \in TP_y \\ \sum_{j_k=1}^s \widehat{a}_{ij_k} \phi_{j_k}(\tau_k) & \textit{if } \tau_k \in TP_z \end{cases} \tag{2.7}$$

for $\tau = [\tau_1, \ldots, \tau_m]_y$ *or* $\tau = [\tau_1, \ldots, \tau_m]_z$.

Proof. These formulas result from Lemma 2.2 by writing $(hk_i, h\ell_i)$ from the formulas (II.2.2) as a P-series $(hk_i, h\ell_i) = P\big(\phi_i, (y_0, z_0)\big)$ so that

$$(h \sum_j a_{ij} k_j, h \sum_j \widehat{a}_{ij} \ell_j) = P\big(\psi_i, (y_0, z_0)\big)$$

is also a P-series. Observe that equation (2.6) corresponds to (1.16) (where $\mathbf{g}_i$ has to be replaced with ϕ_i) and that formula (2.7) comprises (1.13) and (1.15), where we now write ψ_i instead of $\mathbf{u}_i$. □

The expressions $\phi(\tau)$ are shown in Table 2.1 for all trees in TP_y up to order $|\tau| \le 3$. A similar table must be added for trees in TP_z, where all roots are white and all b_i are replaced with $\widehat{b}_i$. The general rule is the following: attach to every vertex a summation index. Then, the expression $\phi(\tau)$ is a sum over all summation indices with the summand being a product of b_i or $\widehat{b}_i$ (depending on whether the root "i" is black or white) and of a_{jk} (if "k" is black) or $\widehat{a}_{jk}$ (if "k" is white), for each vertex "k" directly above "j".

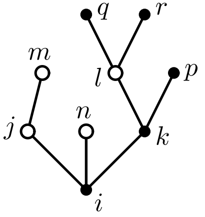
Theorem 2.5 (Order Conditions). *A partitioned Runge–Kutta method (II.2.2) has order r, i.e.,* $y_1 - y(t_0 + h) = \mathcal{O}(h^{r+1})$, $z_1 - z(t_0 + h) = \mathcal{O}(h^{r+1})$, *if and only if*

$$\phi(\tau) = \frac{1}{\gamma(\tau)} \qquad \textit{for } \tau \in TP_y \cup TP_z \textit{ with } |\tau| \le r. \tag{2.8}$$

Proof. This corresponds to Theorem 1.5 and is seen by comparing the expansions of Theorems 2.4 and 2.3. □

Example 2.6. We see that not only does every individual Runge–Kutta method have to be of order r, but also the so-called *coupling conditions* between the coefficients of both methods must hold. The order conditions mentioned above (see formulas (II.2.3) and (II.2.5)) correspond to the trees , , and . For the tree sketched below we obtain

$$\sum_{i,j,k,l,m,n,p,q,r} b_i \widehat{a}_{ij} \widehat{a}_{jm} \widehat{a}_{in} a_{ik} \widehat{a}_{kl} a_{lq} a_{lr} a_{kp} = \frac{1}{9 \cdot 2 \cdot 5 \cdot 3}$$



or, by using $\sum_j a_{ij} = c_i$ and $\sum_j \widehat{a}_{ij} = \widehat{c}_i$,

$$\sum_{i,j,k,l} b_i \widehat{c}_i \widehat{a}_{ij} \widehat{c}_j a_{ik} c_k \widehat{a}_{kl} c_l^2 = \frac{1}{270}.$$

III.2.3 Order Conditions for Nyström Methods

A "modern" order theory for Nyström methods (II.2.11) of Sect. II.2.3 was first given in 1976 by Hairer & Wanner (see Sect. II.14 of Hairer, Nørsett & Wanner

1993). Later it turned out that these conditions are obtained easily by applying the theory of partitioned Runge–Kutta methods to the system

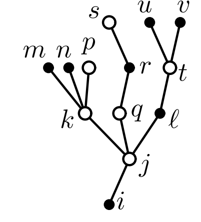
$$\dot{y} = z \qquad \dot{z} = g(y, z), \tag{2.9}$$

which is of the form (2.1). This function has the partial derivative $f_z = I$ and all other derivatives of f are zero. As a consequence, many elementary differentials are zero and the corresponding order conditions can be omitted. The only trees remaining are those for which

"black vertices have at most one son and this son must be white". (2.10)

Example 2.7. The tree sketched below apparently satisfies condition (2.10) and the corresponding order condition becomes, by Theorem 2.4 and formula (2.8),

$$\sum_{i,j,k,\ldots,v} b_i \widehat{a}_{ij} \widehat{a}_{jk} a_{km} a_{kn} \widehat{a}_{kp} \widehat{a}_{jq} a_{qr} \widehat{a}_{rs} a_{j\ell} \widehat{a}_{\ell t} a_{tu} a_{tv} = \frac{1}{13\cdot 12\cdot 4\cdot 3\cdot 2\cdot 4\cdot 3}\,.$$



Due to property (2.10), each a_{ik} inside the tree comes with a corresponding $\widehat{a}_{kj}$, and by (2.10), both factors contract to an $\overline{a}_{ij}$; similarly, the black root is only connected to one white vertex, the corresponding $b_i \widehat{a}_{ij}$ simplifies to $\overline{b}_j$. We thus get

$$\sum_{j,k,q,s,t} \overline{b}_j \widehat{a}_{jk} c_k^2 \widehat{c}_k \widehat{a}_{jq} \overline{a}_{qs} \overline{a}_{jt} c_t^2 = \frac{1}{13 \cdot 3456}.$$

Each of the above order conditions for a tree in TP_y has a "twin" in TP_z of one order lower with the root cut off. For the above example this twin becomes

$$\sum_{j,k,q,s,t} b_j \widehat{a}_{jk} c_k^2 \widehat{c}_k \widehat{a}_{jq} \overline{a}_{qs} \overline{a}_{jt} c_t^2 = \frac{1}{3456}.$$

We need only consider the trees in TP_z if

$$\overline{b}_i = b_i(1 - c_i)$$

is satisfied (see Lemma II.14.13 of Hairer, Nørsett & Wanner (1993), Sect. II.14).

Remark 2.8. Strictly speaking, the theory of partitioned methods is applicable to Nyström methods only if the matrix $(\widehat{a}_{ij})$ is invertible. However, since we arrive at expansions with a finite number of algebraic conditions, we can recover the singular case by a continuous perturbation of the coefficients.

Equations without Friction. Although condition (2.10) already eliminates many order conditions, Nyström methods for the general problem $\ddot{y} = g(y, \dot{y})$ cannot be much better than an excellent Runge–Kutta method applied pairwise to system (2.9).

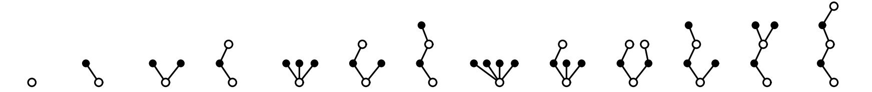
There *is*, however, an important special case where much more progress is possible, namely equations of the type

$$\ddot{y} = g(y), \tag{2.11}$$

which corresponds to motion without friction. In this case, the function for $\dot{z}$ in (2.9) is *independent of* z, and in addition to (2.10) we have a second condition, namely

$$\text{“white vertices have only black sons”}. \tag{2.12}$$

Both conditions reduce the remaining trees drastically. Along each branch, there occur alternating black and white vertices. Ramifications only happen at white vertices. This case allows the construction of excellent numerical methods of high orders. For example, the following 13 trees



assure order 5, whereas ordinary Runge–Kutta theory requires 17 conditions for this order. See Hairer, Nørsett & Wanner (1993), pages 291f, for tables, examples and references.

III.3 Order Conditions for Composition Methods

We have seen in the preceding chapter that composition methods of arbitrarily high order can be obtained with the use of Theorem II.4.1. However, as demonstrated in Fig. II.4.4, these methods are not attractive for high orders. This section is devoted to the derivation of order conditions, which then allow the construction of optimal high order composition methods.

The order conditions for these methods are often derived via the Baker-Campbell-Hausdorff formula. This will be the subject of Sect. III.5 below. Only very recently, Murua & Sanz-Serna (1999) have found an elegant theory based on the idea of B-series. This paper has largely inspired the subsequent presentation.

III.3.1 Introduction

The principal tool in this section is the Taylor series expansion

$$\Phi_h(y) = y + hd_1(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \tag{3.1}$$

of the basic method. The only hypothesis which we require for this method is *consistency*, i.e., that

$$d_1(y) = f(y). \tag{3.2}$$

All other functions $d_i(y)$ are arbitrary.

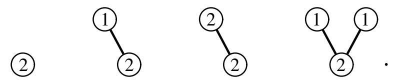
The underlying idea for obtaining the expansions for composition methods is, in fact, very simple: we just insert the series (3.1), with varying values of h, into itself. All our experience from Sect. III.1.2 with the insertion of a B-series into a function will certainly be helpful. We demonstrate this for the case of the composition $\Psi_h = \Phi_{\alpha_2 h} \circ \Phi_{\alpha_1 h}$. Applied to an initial value y_0, this gives with (3.1)

$$\begin{aligned} y_1 &= \Phi_{\alpha_1 h}(y_0) = y_0 + h\alpha_1 d_1(y_0) + h^2\alpha_1^2 d_2(y_0) + \ldots \\ y_2 &= \Phi_{\alpha_2 h}(y_1) = y_1 + h\alpha_2 d_1(y_1) + h^2\alpha_2^2 d_2(y_1) + \ldots . \end{aligned} \tag{3.3}$$

We now insert the first series into the second, in the same way as we did in (1.35). Then, for example, the term $h^2\alpha_2^2 d_2(y_1)$ becomes

$$\begin{aligned} y_2 = \ldots \; &+ \; h^2\alpha_2^2 d_2(y_0) + h^3\alpha_2^2\alpha_1 d_2'(y_0)d_1(y_0) \\ &+ \; h^4\alpha_2^2\alpha_1^2 d_2'(y_0)d_2(y_0) + \frac{h^4}{2}\alpha_2^2\alpha_1^2 d_2''(y_0)(d_1(y_0), d_1(y_0)) + \ldots \end{aligned} \tag{3.4}$$

We see that we arrive at "generalized" B-series, where the elementary differentials contain not only *one* function, but are composed of *infinitely many* functions and their derivatives. We symbolize the four terms written in (3.4) by the trees



This leads us to the following definition.

Definition 3.1 **(∞-Trees, B_∞-series).** We extend Definitions 1.1 and 1.2 to T_∞, the *set of all rooted trees where each vertex bears a positive integer* without any further restriction, and use the notation

$$\begin{aligned} ①, ②, ③, \ldots &= \text{the trees with one vertex;} \\ [\tau_1, \ldots, \tau_m]_i &= \text{the tree } \tau \text{ formed by a new root } ⓘ \text{ connected to } \tau_1, \ldots, \tau_m; \\ F(ⓘ)(y) &= d_i(y); \\ F(\tau)(y) &= d_i^{(m)}(y)(F(\tau_1)(y), \ldots, F(\tau_m)(y)) \;\; \text{for } \tau \text{ as above;} \\ |\tau| &= 1 + |\tau_1| + \ldots + |\tau_m|, \quad \text{the number of vertices of } \tau; \\ ||\tau|| &= i + ||\tau_1|| + \ldots + ||\tau_m||, \quad \text{the sum of the labels of } \tau; \\ \sigma(\tau) &= \mu_1!\mu_2! \cdot \ldots \cdot \sigma(\tau_1) \cdot \ldots \cdot \sigma(\tau_m), \\ &\quad \text{where } \mu_1, \mu_2, \ldots \text{ count equal trees among } \tau_1, \ldots, \tau_m, \\ &\quad \text{the } \textit{symmetry coefficient} \text{ respecting the labels;} \\ i(\tau) &= i, \quad \text{the } \textit{label of the root} \text{ of } \tau. \end{aligned}$$

For a map $a : T_\infty \cup \{\emptyset\} \to \mathbb{R}$ we write

$$B_\infty(a, y) = a(\emptyset)y + \sum_{\tau \in T_\infty} \frac{h^{||\tau||}}{\sigma(\tau)}\, a(\tau)\, F(\tau)(y) \tag{3.5}$$

which extends the *notion of B-series* to the new situation.

Example 3.2. For the tree

$$\tau = [\text{①}, [\text{⑤},\text{⑥},\text{⑥}]_7]_4 \quad \Leftrightarrow \quad \tau = [\tau_1, \tau_2]_4 \quad \text{where} \quad \tau_1 = \text{①}, \;\; \tau_2 = [\text{⑤},\text{⑥},\text{⑥}]_7 \tag{3.6}$$

we have

$$F(\tau)(y) = d_4''(y)\big(d_1(y), d_7'''(y)\big(d_5(y), d_6(y), d_6(y)\big)\big)$$

$$\tau = [\text{①}, [\text{⑤},\text{⑥},\text{⑥}]_7]_4, \quad |\tau| = 6, \;\; ||\tau|| = 29, \;\; \sigma(\tau) = 2, \;\; i(\tau) = 4\,.$$

The above calculations for (3.4) are governed by the following lemma.

Lemma 3.3. *For a series $B_\infty(a, y)$ with $a(\emptyset) = 1$ we have*

$$h^i d_i\Big(B_\infty(a, y)\Big) = \sum_{\tau \in T_\infty, i(\tau) = i} \frac{h^{||\tau||}}{\sigma(\tau)}\, a'(\tau)\, F(\tau)(y), \tag{3.7}$$

where $a'(\text{ⓘ}) = 1$ *and*

$$a'(\tau) = a(\tau_1) \cdot \ldots \cdot a(\tau_m) \qquad \textit{for } \; \tau = [\tau_1, \ldots, \tau_m]_i. \tag{3.8}$$

Proof. This is a straightforward extension of Lemma 1.9 with exactly the same proof. □

The preceding lemma leads directly to the order conditions for composition methods. However, if we continue with compositions of the type (II.4.1), we arrive at conditions without real solutions. We therefore turn to compositions including the adjoint method as well.

III.3.2 The General Case

As in (II.4.6), we consider

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi^*_{\beta_s h} \circ \ldots \circ \Phi_{\alpha_2 h} \circ \Phi^*_{\beta_2 h} \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}, \tag{3.9}$$

and we obtain with the help of the above lemma the corresponding B_∞-series.

Lemma 3.4 (Recurrence Relations). *The following compositions are B_∞-series*

$$\begin{aligned} \big(\Phi^*_{\beta_k h} \circ \ldots \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}\big)(y) &= B_\infty(b_k, y) \\ \big(\Phi_{\alpha_k h} \circ \Phi^*_{\beta_k h} \circ \ldots \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}\big)(y) &= B_\infty(a_k, y). \end{aligned} \tag{3.10}$$

Their coefficients are recursively given by $a_k(\emptyset) = 1$, $b_k(\emptyset) = 1$, $a_0(\tau) = 0$ for all $\tau \in T_\infty$, and

$$\begin{aligned} b_k(\tau) &= a_{k-1}(\tau) - (-\beta_k)^{i(\tau)}\, b_k'(\tau), \\ a_k(\tau) &= b_k(\tau) + \alpha_k^{i(\tau)}\, b_k'(\tau). \end{aligned} \tag{3.11}$$

Proof. The coefficients $a_0(\tau)$ correspond to the identity map $B_\infty(a_0, y) = y$. The second formula of (3.11) follows from

$$B_\infty(a_k, y) = \Phi_{\alpha_k h}\Big(B_\infty(b_k, y)\Big) = B_\infty(b_k, y) + \sum_{i\ge 1} \alpha_k^i h^i d_i\Big(B_\infty(b_k, y)\Big),$$

and from an application of Lemma 3.3.

The relation $B_\infty(b_k, y) = \Phi^*_{\beta_k h}\big(B_\infty(a_{k-1}, y)\big)$, which involves the adjoint method, needs a little trick: we write it as $B_\infty(a_{k-1}, y) = \Phi_{-\beta_k h}\big(B_\infty(b_k, y)\big)$ (remember that $\Phi^*_h = \Phi^{-1}_{-h}$), apply Lemma 3.3 again, and reverse the formula. This gives the first equation of (3.11). □

Adding the equations of (3.11), we get

$$a_k(\tau) = a_{k-1}(\tau) + \big(\alpha_k^{i(\tau)} - (-\beta_k)^{i(\tau)}\big) b_k'(\tau). \tag{3.12}$$

Because of $b_k'(\textcircled{i}) = 1$, we obtain

$$\begin{aligned} a_k(\textcircled{i}) &= \sum_{\ell=1}^{k} \big(\alpha_\ell^i - (-\beta_\ell)^i\big) \\ b_k(\textcircled{i}) &= \sum_{\ell=1}^{k-1} \alpha_\ell^i - \sum_{\ell=1}^{k} (-\beta_\ell)^i \;=\; {\sum_{\ell=1}^{k}}' \big(\alpha_\ell^i - (-\beta_\ell)^i\big). \end{aligned} \tag{3.13}$$

The fact that, for $b_k(\textcircled{i})$, the sum of $(-\beta_\ell)^i$ is from 1 to k, but the sum of α_ℓ^i is only from 1 to $k-1$, has been *indicated by a prime* attached to the summation symbol. Continuing to apply the formulas (3.11) and (3.12) to more and more complicated trees, we quickly understand the general rule for the coefficients of an arbitrary tree.

Example 3.5. The tree τ in (3.6) gives

$$\begin{aligned} a_s(\tau) = &\sum_{k=1}^{s} (\alpha_k^4 - \beta_k^4) {\sum_{\ell=1}^{k}}' (\alpha_\ell + \beta_\ell) \\ &\cdot {\sum_{m=1}^{k}}' (\alpha_m^7 + \beta_m^7) {\sum_{n=1}^{m}}' (\alpha_n^5 + \beta_n^5) \Big({\sum_{p=1}^{m}}' (\alpha_p^6 - \beta_p^6)\Big)^2 . \end{aligned} \tag{3.14}$$

The Order Conditions. The exact solution of $\dot y = f(y)$ is a B-series $y(t_0 + h) = B(\mathbf{e}, y_0)$ (see (1.26)). Since $d_1(y) = f(y)$, every B-series is also a B_∞-series with $\mathbf{e}(\tau) = 0$ for trees with at least one label different from 1. Therefore, we also have $y(t_0 + h) = B_\infty(\mathbf{e}, y_0)$, where the coefficients $\mathbf{e}(\tau)$ satisfy $\mathbf{e}(\textcircled{1}) = 1$, $\mathbf{e}(\tau) = 0$ if $i(\tau) > 1$, and

$$\mathbf{e}(\tau) = \frac{1}{|\tau|}\, \mathbf{e}(\tau_1) \cdot \ldots \cdot \mathbf{e}(\tau_m) \qquad \text{for } \tau = [\tau_1, \ldots, \tau_m]_1. \tag{3.15}$$

Theorem 3.6. *The composition method* $\Psi_h(y) = B_\infty(a_s, y)$ *of (3.9) has order* p *if*

$$a_s(\tau) = \mathbf{e}(\tau) \qquad \textit{for } \tau \in T_\infty \textit{ with } ||\tau|| \le p. \tag{3.16}$$

Proof. This follows from a comparison of the B_∞-series for the numerical and the exact solution. For the necessity of (3.16), the independence of the elementary differentials has to be studied as in Exercise 3. □

III.3.3 Reduction of the Order Conditions

The order conditions of the foregoing section are indeed beautiful, but for the moment they are not of much use, because of the enormous number of trees in T_∞ of a certain order. For example, there are 166 trees in T_∞ with $||\tau|| \le 6$. Fortunately, the equations are not all independent, as we shall see now.

Definition 3.7 (Butcher 1972, Murua & Sanz-Serna 1999). For two trees in T_∞, $u = [u_1, \ldots, u_m]_i$ and $v = [v_1, \ldots, v_l]_j$, we denote

$$u \circ v := [u_1, \ldots, u_m, v]_i \, , \qquad u \times v := [u_1, \ldots, u_m, v_1, \ldots, v_l]_{i+j} \tag{3.17}$$

and call them the *Butcher product* and *merging product*, respectively (see Fig. 3.1).

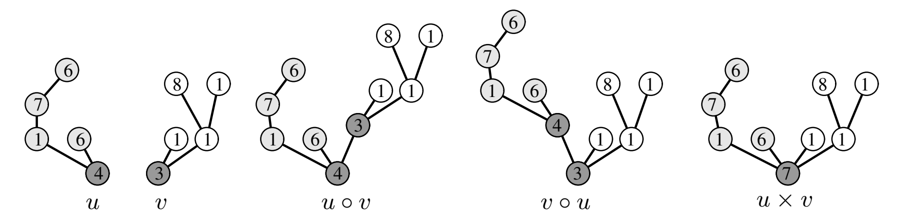


Fig. 3.1. The Butcher product and the merging product

The merging product is associative and commutative, the Butcher product is neither of the two. To simplify the notation, we write products of *several* factors without parentheses, when we mean evaluation from left to right:

$$u \circ v_1 \circ v_2 \circ \ldots \circ v_s = (((u \circ v_1) \circ v_2) \circ \ldots) \circ v_s. \tag{3.18}$$

Here the factors $v_1, \ldots, v_s$ can be freely permuted.

All subsequent results concern properties of $a_k(\tau)$ as well as $b_k(\tau)$, valid for all k. To avoid writing all formulas twice, we replace $a_k(\tau)$ and $b_k(\tau)$ everywhere by a neutral symbol $c(\tau)$.

Lemma 3.8 (Switching Lemma). *All a_k, b_k of Lemma 3.4 satisfy, for all $u, v \in T_\infty$, the relation*

$$c(u \circ v) + c(v \circ u) = c(u) \cdot c(v) - c(u \times v). \tag{3.19}$$

Proof. The recursion formulas (3.11) are of the form

$$a(\tau) = b(\tau) + \alpha^{i(\tau)}\, b'(\tau). \tag{3.20}$$

We arrange this formula, for all five trees of Fig. 3.1, as follows:

$$\begin{array}{ll} & a(u \circ v) \quad + \quad a(v \circ u) \quad + \quad a(u \times v) \quad - \quad a(u)a(v) \\ = & b(u \circ v) \quad + \quad b(v \circ u) \quad + \quad b(u \times v) \quad - \quad b(u)b(v) \\ + & \alpha^{i(u)} b'(u \circ v) + \alpha^{i(v)} b'(v \circ u) + \alpha^{i(u)+i(v)} b'(u \times v) \\ - & \alpha^{i(u)} b'(u) b(v) - \alpha^{i(v)} b'(v) b(u) - \alpha^{i(u)} \alpha^{i(v)} b'(u) b'(v)\ . \end{array}$$

Because of $b'(u \circ v) = b'(u)b(v)$ and $b'(u \times v) = b'(u)b'(v)$, the last two rows cancel, hence

$$a(\tau) \ \text{ satisfies (3.19)} \quad \Leftrightarrow \quad b(\tau) \ \text{ satisfies (3.19)}. \tag{3.21}$$

Thus, beginning with a_0, then b_1, then a_1, etc., all a_k and b_k must satisfy (3.19). □

The Switching Lemma 3.8 reduces considerably the number of order conditions. Since the right-hand expression involves only trees with $|\tau| < |u \circ v|$, and since relation (3.19) is also satisfied by $\mathbf{e}(\tau)$, an induction argument shows that the order conditions (3.16) for the trees $u \circ v$ and $v \circ u$ are equivalent. The operation $u \circ v \mapsto v \circ u$ consists simply in switching the root from one vertex to the next. By repeating this argument, we see that we can freely move the root inside the graph, and of all these trees, only one needs to be retained. For order 6, for example, there remain 68 conditions out of the original 166.

Our next results show how relation (3.19) also generates a considerable amount of reductions of the order conditions. These ideas (for the special situation of symplectic methods) have already been exploited by Calvo & Hairer (1995b).

Lemma 3.9. *Assume that all b_k of Lemma 3.4 satisfy a relation of the form*

$$\sum_{i=1}^{N} A_i \prod_{j=1}^{m_i} c(u_{ij}) = 0 \tag{3.22}$$

with all $m_i > 0$. Then, for any tree w, all a_k and b_k satisfy the relation

$$\sum_{i=1}^{N} A_i\, c(w \circ u_{i1} \circ u_{i2} \circ \ldots \circ u_{i,m_i}) = 0. \tag{3.23}$$

Proof. The relation (3.20), written for the tree $w \circ u_{i1} \circ u_{i2} \circ \ldots \circ u_{i,m_i}$, is

$$\begin{aligned} a(w \circ u_{i1} \circ \ldots \circ u_{i,m_i}) &= b(w \circ u_{i1} \circ \ldots \circ u_{i,m_i}) \\ &+ \alpha^{i(w)} \, b'(w) b(u_{i1}) \cdot \ldots \cdot b(u_{i,m_i}). \end{aligned}$$

Multiplying with A_i and summing over i, this shows that, under the hypothesis (3.22) for b, the relation (3.23) holds for b if and only if it holds for a. The coefficients $a_0(\tau) = 0$ for the identity map satisfy (3.22) and (3.23) because $m_i > 0$. Starting from this, we again conclude (3.23) recursively for all a_k and b_k. □

The following lemma [3] extends formula (3.19) to the case of *several* factors.

Lemma 3.10. *For any three trees u, v, w all a_k, b_k of Lemma 3.4 satisfy a relation*

$$c(u \circ v \circ w) + c(v \circ u \circ w) + c(w \circ u \circ v) = c(u) \cdot c(v) \cdot c(w) + \ldots\ , \tag{3.24}$$

where the dots indicate a linear combination of products $\prod_j c(v_j)$ with $|v_1| + |v_2| + \ldots < |u| + |v| + |w|$ and, for each term, at least one of the v_j possesses a label larger than one. The general formula, for m trees $u_1, \ldots, u_m$, is

$$\sum_{i=1}^{m} c\big(u_i \circ u_1 \circ \ldots \circ u_{i-1} \circ u_{i+1} \circ \ldots \circ u_m\big) = \prod_{i=1}^{m} c(u_i) + \ldots\ . \tag{3.25}$$

Proof. We apply Lemma 3.9 to (3.19) and obtain

$$c(w \circ (u \circ v)) + c(w \circ (v \circ u)) = c(w \circ u \circ v) - c(w \circ (u \times v)). \tag{3.26}$$

Next, we apply the Switching Lemma 3.8 to the trees to the left and get

$$\begin{aligned} c(w \circ (u \circ v)) + c(u \circ v \circ w) &= c(w) \cdot c(u \circ v) - c(w \times (u \circ v)) \\ c(w \circ (v \circ u)) + c(v \circ u \circ w) &= c(w) \cdot c(v \circ u) - c(w \times (v \circ u)). \end{aligned}$$

Adding these formulas and subtracting (3.26) gives

$$c(u \circ v \circ w) + c(v \circ u \circ w) + c(w \circ u \circ v) = c(w)\big(c(u \circ v) + c(v \circ u)\big) + \ldots$$

which becomes (3.24) after another use of the Switching Lemma. Thereby, everything which goes into "$+\ldots$" contains somewhere a merging product, whose roots introduce necessarily labels larger than one.

Continuing like this, we get recursively (3.25) for all m. □

In order that the further simplifications do not turn into chaos, we fix, once and for all, a *total order relation* (written $<$) on T_∞, where we only require that the order respects the number of vertices, i.e., that

$$u < v \quad \text{whenever} \quad |u| < |v|. \tag{3.27}$$

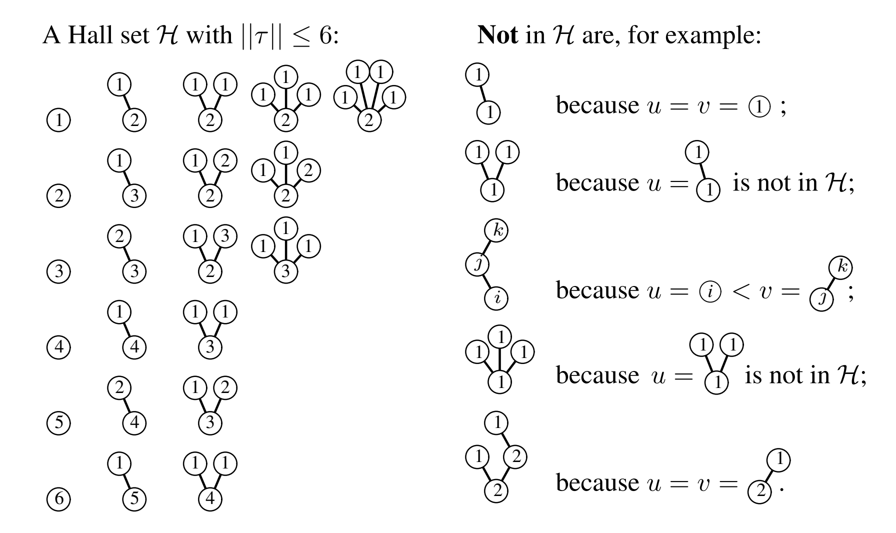
Similar to the strategy introduced by Hall (1950) for simplifying bracket expressions in Lie algebras, we define the following subset of T_∞.

[3] due to A. Murua, private communication, Feb. 2001

Definition 3.11 (Hall Set). The *Hall set* corresponding to an order relation (3.27) is a subset $\mathcal{H} \subset T_\infty$ defined by

$$\begin{array}{lll} ⓘ \in \mathcal{H} & \text{for } i = 1, 2, 3, \ldots & \\ \tau \in \mathcal{H} & \Leftrightarrow \quad \text{there exist } u, v \in \mathcal{H},\, u > v, \text{ such that} & \tau = u \circ v. \end{array}$$

Example 3.12. The trees in the subsequent table are ordered from left to right with respect to $|\tau|$, and from top to bottom within fixed $|\tau|$. There remain finally 22 conditions for order 6.



Theorem 3.13 (Murua & Sanz-Serna 1999). *For each $\tau \in T_\infty$ there are constants A_i, integers m_i and trees $u_{ij} \in \mathcal{H}$ such that for all a_k, b_k of Lemma 3.4 we have*

$$c(\tau) = \sum_{i=1}^{N} A_i \prod_{j=1}^{m_i} c(u_{ij}), \qquad u_{ij} \in \mathcal{H},\ |u_{i1}| + \ldots + |u_{i,m_i}| \leq |\tau|. \tag{3.28}$$

Proof. We proceed by induction on $|\tau|$. For $\tau =$ ⓘ the statement is trivial, because ⓘ $\in \mathcal{H}$. We thus consider $\tau \in T_\infty$ with $|\tau| \geq 2$, write it as $\tau = u \circ v$, and conclude through the following two steps.

First Step. We apply the induction hypothesis (3.28) to v, i.e.,

$$c(v) = \sum_i B_i \prod_j c(v_{ij}), \qquad v_{ij} \in \mathcal{H}, \quad \textstyle\sum_j |v_{ij}| \leq |v|. \tag{3.29}$$

To this, we apply Lemma 3.9 followed by the Switching Lemma 3.8:

$$\begin{aligned} c(\tau) = c(u \circ v) &= \sum_i B_i\, c(u \circ v_{i1} \circ v_{i2} \ldots \circ v_{i,n_i}) \\ &= -\sum_i B_i\, c\big(v_{in_i} \circ (u \circ v_{i1} \circ \ldots \circ v_{i,n_i-1})\big) + \ldots . \end{aligned}$$

The "$+\ldots$" indicate terms containing trees to which we can apply our induction hypothesis. Inside the above expressions, we apply the induction hypothesis to the trees $u \circ v_{i1} \circ \ldots \circ v_{i,n_i-1}$, followed once again by Lemma 3.9. We arrive at a huge double sum which constitutes a linear combination of expressions of the form

$$c\big(u_1 \circ u_2 \circ \ldots \circ u_m\big) \tag{3.30}$$

and of terms "$+\ldots$" covered by the induction hypothesis. The point of the above dodges was *to make sure that all* $u_1, u_2, \ldots, u_m$ *are in* $\mathcal{H}$.

Second Step. It remains to reduce an expression (3.30) to the form required by (3.28). The trees $u_2, \ldots, u_m$ can be permuted arbitrarily; we arrange them in increasing order $u_2 \leq \ldots \leq u_m$.

Case 1. If $u_1 > u_2$, then by definition $u_1 \circ u_2 = w \in \mathcal{H}$ and we absorb the second factor into the first and obtain a product $w \circ u_3 \circ \ldots \circ u_m$ with *fewer* factors.

Case 2. If $u_1 < u_2 \leq \ldots$, we shuffle the factors with the help of Lemma 3.10 and obtain for (3.30) the expression

$$-\sum_{i=2}^{m} c(u_i \circ u_1 \circ \ldots) + \prod_{i=1}^{m} c(u_i) + \ldots\ .$$

With the first terms we return to Case 1, the second term is precisely as in (3.28), and the terms "$+\ldots$" are covered by the induction hypothesis.

Case 3. Now let $u_1 = u_2 < \ldots$. In this case, the formula (3.25) of Lemma 3.10 contains the term (3.30) twice. We group both together, so that (3.30) becomes

$$-\frac{1}{2}\sum_{i=3}^{m} c(u_i \circ u_1 \circ u_1 \circ \ldots) + \frac{1}{2}\prod_{i=1}^{m} c(u_i) + \ldots$$

and we go back to Case 1. If the first *three* trees are equal, we group three equal terms together and so on.

The whole reduction process is repeated until all Butcher products have disappeared. □

Theorem 3.14 (Murua & Sanz-Serna 1999). *The composition method* $\Psi_h(y) = B_\infty(a_s, y)$ *of (3.9) has order* p *if and only if*

$$a_s(\tau) = \mathbf{e}(\tau) \qquad \textit{for } \ \tau \in \mathcal{H} \textit{ with } \|\tau\| \leq p.$$

The coefficients $\mathbf{e}(\tau)$ *are those of Theorem 3.6.*

Proof. We have seen in Sect. II.4 that composition methods of arbitrarily high order exist. Since the coefficients A_i of (3.28) do not depend on the mapping $c(\tau)$, this together with Theorem 3.6 implies that the relation (3.28) is also satisfied by the mapping $\mathbf{e}$ for the exact solution. This proves the statement. □

Example 3.15. The order conditions for orders $p = 1, \ldots, 4$ become, with the trees of Example 3.12 and the rule of (3.14), as follows:

Order 1: ① $$\sum_{k=1}^{s}(\alpha_k + \beta_k) = 1$$

Order 2: ② $$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2) = 0$$

Order 3: ③ $$\sum_{k=1}^{s}(\alpha_k^3 + \beta_k^3) = 0$$

①–② $$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2)\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell) = 0 \tag{3.31}$$

Order 4: ④ $$\sum_{k=1}^{s}(\alpha_k^4 - \beta_k^4) = 0$$

①–③ $$\sum_{k=1}^{s}(\alpha_k^3 + \beta_k^3)\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell) = 0$$

① ①–② $$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2)\Big(\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell)\Big)^2 = 0,$$

where, as above, a *prime* attached to a summation symbol indicates that the sum of α_ℓ^i is only from 1 to $k-1$, whereas the sum of $(-\beta_\ell)^i$ is from 1 to k. Similarly, the remaining trees of Example 3.12 with $||\tau|| = 5$ and $||\tau|| = 6$ give the additional conditions for order 5 and 6.

We shall see in Sect. V.3 how further reductions and numerical values are obtained under various assumptions of symmetry.

III.3.4 Order Conditions for Splitting Methods

Splitting methods, introduced in Sect. II.5, are based on differential equations of the form

$$\dot y = f_1(y) + f_2(y), \tag{3.32}$$

where the flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems $\dot y = f_1(y)$ and $\dot y = f_2(y)$ are assumed to be known exactly. In this situation, the method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$$

is of first order and, together with its adjoint $\Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]}$, can be used as the basic method in the composition (3.9). This yields

$$\Psi_h = \varphi_{a_{s+1}h}^{[1]} \circ \varphi_{b_s h}^{[2]} \circ \varphi_{a_s h}^{[1]} \circ \ldots \circ \varphi_{b_2 h}^{[2]} \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \tag{3.33}$$

where

$$b_i = \alpha_i + \beta_i, \qquad a_i = \alpha_{i-1} + \beta_i \tag{3.34}$$

with the conventions $\alpha_0 = 0$ and $\beta_{s+1} = 0$. Consequently, the splitting method (3.33) is a special case of (3.9) and we have the following obvious result.

Theorem 3.16. *Suppose that the composition method (3.9) is of order p for all basic methods Φ_h, then the splitting method (3.33) with a_i, b_i given by (3.34) is of the same order p.* □

We now want to establish the reciprocal result. To every consistent splitting method (3.33), i.e., with coefficients satisfying $\sum_i a_i = \sum_i b_i = 1$, there exist unique α_i, β_i such that (3.34) holds. Does the corresponding composition method have the same order?

Theorem 3.17. *If a consistent splitting method (3.33) is of order p at least for problems of the form (3.32) with the integrable splitting*

$$f_1(y) = \begin{pmatrix} g_1(y_2) \\ 0 \end{pmatrix}, \quad f_2(y) = \begin{pmatrix} 0 \\ g_2(y_1) \end{pmatrix} \quad \textit{where} \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \tag{3.35}$$

then the corresponding composition method has the same order p for an arbitrary basic method Φ_h.

Proof. McLachlan (1995) proves this result in the setting of Lie algebras. We give here a proof using the tools of this section.

a) The flows corresponding to the two vector fields f_1 and f_2 of (3.35) are $\varphi_t^{[1]}(y) = y + tf_1(y)$ and $\varphi_t^{[2]}(y) = y + tf_2(y)$, respectively. Consequently, the method $\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$ can be written in the form (3.1) with

$$d_1(y) = f_1(y) + f_2(y), \qquad d_{k+1}(y) = \frac{1}{k!}\, f_1^{(k)}(y)\Big(f_2(y), \ldots, f_2(y)\Big). \tag{3.36}$$

The idea is to construct, for every tree $\tau \in \mathcal{H}$, functions $g_1(y_2)$ and $g_2(y_1)$ such that the first component of $F(\tau)(0)$ is non-zero whereas the first component of $F(\sigma)(0)$ vanishes for all $\sigma \in T_\infty$ different from τ. This construction will be explained in part (b) below. Since the local error of the composition method is a B_∞-series with coefficients $a_s(\tau) - \mathbf{e}(\tau)$, this implies that the order conditions for $\tau \in \mathcal{H}$ with $\|\tau\| \le p$ are necessary already for this very special class of problems. Theorem 3.14 thus proves the statement.

b) For the construction of the functions $g_1(y_2)$ and $g_2(y_1)$ we have to understand the structure of $F(\tau)(y)$ with $d_k(y)$ given by (3.36). Consider for example the tree $\tau \in T_\infty$ of Fig. 3.2, for which we have $F(\tau)(y) = d_2''(y)\big(d_1(y), d_3(y)\big)$. Inserting $d_k(y)$ from (3.36), we get by Leibniz' rule a linear combination of eight expressions ($i \in \{1,2\}$)

$$f_1'''\big(f_2, f_i, f_1''(f_2,f_2)\big), \qquad f_1''\big(f_2'f_i, f_1''(f_2,f_2)\big),$$
$$f_1''\big(f_i, f_2'f_1''(f_2,f_2)\big), \qquad f_1'f_2''\big(f_i, f_1''(f_2,f_2)\big),$$

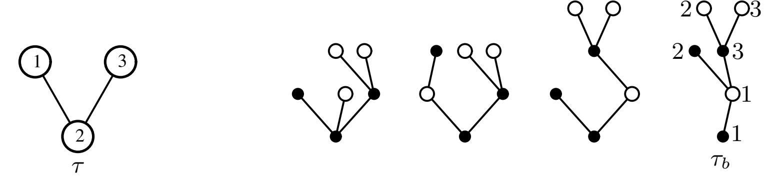


Fig. 3.2. Trees for illustrating the equivalence of the order conditions between composition and splitting methods

each of which can be identified with a bi-coloured tree (see Sect. III.2.1, a vertex $\bullet$ corresponds to f_1 and $\circ$ to f_2). The trees corresponding to these expressions with $i = 1$ are shown in Fig. 3.2. Due to the special form of $d_k(y)$ in (3.36) and due to the fact that in trees of the Hall set $\mathcal{H}$ the vertex ① can appear only at the end of a branch, there is always at least one bi-coloured tree where the vertices $\bullet$ are separated by those of $\circ$ and vice versa. We now select such a tree, denoted by τ_b, and we label the black and white vertices with $\{1, 2, \ldots\}$. We then let $y_1 = (y_1^1, \ldots, y_n^1)^T$ and $y_2 = (y_1^2, \ldots, y_m^2)^T$, where n and m are the numbers of vertices $\bullet$ and $\circ$ in τ_b, respectively. Inspired by "Exercise 4" of Hairer, Nørsett & Wanner (1993), page 155, we define the ith component of $g_1(y_2)$ as the product of all y_j^2 where j runs through the labels of the vertices directly above the vertex $\bullet$ with label i. The function $g_2(y_1)$ is defined similarly. For the example of Fig. 3.2, the tree τ_b yields

$$g_1(y_2) = \begin{pmatrix} y_1^2 \\ y_2^2 y_3^2 \\ 1 \end{pmatrix}, \qquad g_2(y_1) = \begin{pmatrix} y_2^1 y_3^1 \\ 1 \\ 1 \end{pmatrix}.$$

One can check that with this construction the bi-coloured tree τ_b is the only one for which the first component of the elementary differential evaluated at $y = 0$ is different from zero. This in turn implies that among all trees of T_∞ only the tree τ has a non-vanishing first component in its elementary differential. □

Necessity of Negative Steps for Higher Order. One notices that all the composition methods (II.4.6) of oder higher than two with Φ_h given by (II.5.7) lead to a splitting (II.5.6) where at least one of the coefficients a_i and b_i is negative. This may be undesirable, especially when the flow $\varphi_t^{[i]}$ originates from a partial differential equation that is ill-posed for negative time progression. The following result has been proved independently by Sheng (1989) and Suzuki (1991) (see also Goldman & Kaper (1996)). We present the elegant proof found by Blanes & Casas (2005).

Theorem 3.18. *If the splitting method (II.5.6) is of order $p \geq 3$ for general $f^{[1]}$ and $f^{[2]}$, then at least one of the a_i and at least one of the b_i are strictly negative.*

Proof. The condition in equation (3.31) for the tree ③ reads

$$\sum_{k=1}^{s} (\alpha_k^3 + \beta_k^3) = 0 \qquad \text{or also} \qquad \sum_{k=1}^{s+1} (\alpha_{k-1}^3 + \beta_k^3) = 0$$

(remember that $\alpha_0 = 0$ and $\beta_{s+1} = 0$). Now apply the fact that $x^3 + y^3 < 0$ implies $x + y < 0$ and conclude with formulas (3.34). □

III.4 The Baker-Campbell-Hausdorff Formula

This section treats the Baker-Campbell-Hausdorff (short BCH or CBH) formula on the composition of exponentials. It was proposed in 1898 by J.E. Campbell and proved independently by Baker (1905) and Hausdorff (1906). This formula will provide an alternative approach to the order conditions of composition (Sect. II.4) and splitting methods (Sect. II.5). For its derivation we shall use the inverse of the derivative of the exponential function.

III.4.1 Derivative of the Exponential and Its Inverse

Elegant formulas for the derivative of $\exp$ and for its inverse can be obtained by the use of matrix commutators $[\Omega, A] = \Omega A - A\Omega$. If we suppose Ω fixed, this expression defines a linear operator $A \mapsto [\Omega, A]$

$$\mathrm{ad}_{\Omega}(A) = [\Omega, A], \tag{4.1}$$

which is called the adjoint operator (see Varadarajan (1974), Sect. 2.13). Let us start by computing the derivatives of Ω^k. The product rule for differentiation becomes

$$\Big(\frac{d}{d\Omega}\,\Omega^k\Big)H = H\Omega^{k-1} + \Omega H \Omega^{k-2} + \ldots + \Omega^{k-1}H, \tag{4.2}$$

and this equals $kH\Omega^{k-1}$ if Ω and H commute. Therefore, it is natural to write (4.2) as $kH\Omega^{k-1}$ to which are added correction terms involving commutators and iterated commutators. In the cases $k = 2$ and $k = 3$ we have

$$\begin{aligned} H\Omega + \Omega H &= 2H\Omega + \mathrm{ad}_{\Omega}(H) \\ H\Omega^2 + \Omega H \Omega + \Omega^2 H &= 3H\Omega^2 + 3\big(\mathrm{ad}_{\Omega}(H)\big)\Omega + \mathrm{ad}_{\Omega}^{2}(H), \end{aligned}$$

where ad_{Ω}^{i} denotes the iterated application of the linear operator ad_{Ω}. With the convention $\mathrm{ad}_{\Omega}^{0}(H) = H$ we obtain by induction on k that

$$\Big(\frac{d}{d\Omega}\,\Omega^k\Big)H = \sum_{i=0}^{k-1}\binom{k}{i+1}\Big(\mathrm{ad}_{\Omega}^{i}(H)\Big)\Omega^{k-i-1}. \tag{4.3}$$

This is seen by applying Leibniz' rule to $\Omega^{k+1} = \Omega \cdot \Omega^k$ and by using the identity $\Omega\big(\mathrm{ad}_{\Omega}^{i}(H)\big) = \big(\mathrm{ad}_{\Omega}^{i}(H)\big)\Omega + \mathrm{ad}_{\Omega}^{i+1}(H)$.

Lemma 4.1. *The derivative of* $\exp\Omega = \sum_{k\ge 0}\frac{1}{k!}\,\Omega^k$ *is given by*

$$\Big(\frac{d}{d\Omega}\,\exp\Omega\Big)H = \Big(d\exp_{\Omega}(H)\Big)\exp\Omega,$$

where

$$d\exp_{\Omega}(H) = \sum_{k\ge 0}\frac{1}{(k+1)!}\,\mathrm{ad}_{\Omega}^{k}(H). \tag{4.4}$$

The series (4.4) converges for all matrices Ω.

Proof. Multiplying (4.3) by $(k!)^{-1}$ and summing, then exchanging the sums and putting $j = k - i - 1$ yields

$$\begin{aligned}\Big(\frac{d}{d\Omega}\exp\Omega\Big)H &= \sum_{k\ge 0}\frac{1}{k!}\sum_{i=0}^{k-1}\binom{k}{i+1}\Big(\mathrm{ad}_{\Omega}^{\,i}(H)\Big)\Omega^{k-i-1}\\ &= \sum_{i\ge 0}\sum_{j\ge 0}\frac{1}{(i+1)!\,j!}\Big(\mathrm{ad}_{\Omega}^{\,i}(H)\Big)\Omega^{j}.\end{aligned}$$

The convergence of the series follows from the boundedness of the linear operator ad_{Ω} (we have $\|\mathrm{ad}_{\Omega}\| \le 2\|\Omega\|$). □

Lemma 4.2 (Baker 1905). *If the eigenvalues of the linear operator* ad_{Ω} *are different from* $2\ell\pi i$ *with* $\ell \in \{\pm 1, \pm 2, \ldots\}$*, then* $d\exp_{\Omega}$ *is invertible. Furthermore, we have for* $\|\Omega\| < \pi$ *that*

$$d\exp_{\Omega}^{-1}(H) = \sum_{k\ge 0}\frac{B_k}{k!}\,\mathrm{ad}_{\Omega}^{\,k}(H), \tag{4.5}$$

where B_k *are the Bernoulli numbers, defined by* $\sum_{k\ge 0}(B_k/k!)x^k = x/(e^x - 1)$.

Proof. The eigenvalues of $d\exp_{\Omega}$ are $\mu = \sum_{k\ge 0}\lambda^k/(k+1)! = (e^{\lambda} - 1)/\lambda$, where λ is an eigenvalue of ad_{Ω}. By our assumption, the values μ are non-zero, so that $d\exp_{\Omega}$ is invertible. By definition of the Bernoulli numbers, the composition of (4.5) with (4.4) gives the identity. Convergence for $\|\Omega\| < \pi$ follows from $\|\mathrm{ad}_{\Omega}\| \le 2\|\Omega\|$ and from the fact that the radius of convergence of the series for $x/(e^x - 1)$ is 2π. □

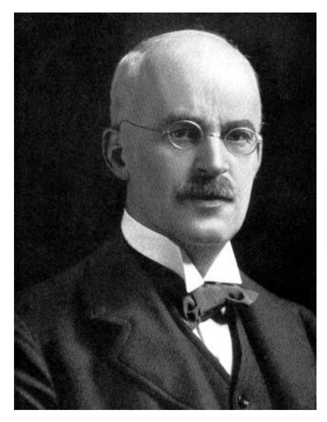
III.4.2 The BCH Formula

Let A and B be two arbitrary (in general non-commuting) matrices. The problem is to find a matrix $C(t)$, such that

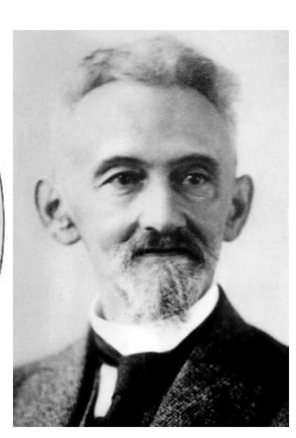
$$\exp(tA)\exp(tB) = \exp C(t). \tag{4.6}$$

In order to get a first idea of the form of $C(t)$, we develop the expression to the left in a series: $\exp(tA)\exp(tB) = I + t(A+B) + \frac{t^2}{2}(A^2 + 2AB + B^2) + \mathcal{O}(t^3) =: I + X$. For sufficiently small t (hence $\|X\|$ is small), the series expansion of the logarithm $\log(I+X) = X - X^2/2 + \ldots$ yields a matrix $C(t) = \log(I+X) = t(A+B) + \frac{t^2}{2}\big(A^2 + 2AB + B^2 - (A+B)^2\big) + \mathcal{O}(t^3)$, which satisfies (4.6). This series has a positive radius of convergence, because it is obtained by elementary operations of convergent series.

The main problem of the derivation of the BCH formula is to get explicit formulas for the coefficients of the series for $C(t)$, and to express the coefficients of $t^2, t^3, \ldots$ in terms of commutators. With the help of the following lemma, recurrence relations for these coefficients will be obtained, which allow for an easy computation of the first terms.


John Edward Campbell [4]


Henry Frederick Baker [5]


Felix Hausdorff [6]

Lemma 4.3. *Let A and B be (non-commuting) matrices. Then, (4.6) holds, where $C(t)$ is the solution of the differential equation*

$$\dot{C} = A + B + \frac{1}{2}\,[A - B, C] + \sum_{k\geq 2} \frac{B_k}{k!}\,\mathrm{ad}\,_C^k(A + B) \tag{4.7}$$

with initial value $C(0) = 0$. Recall that $\mathrm{ad}\,_C A = [C, A] = CA - AC$, and that B_k denote the Bernoulli numbers as in Lemma 4.2.

Proof. We follow Varadarajan (1974), Sect. 2.15, and we consider for small s and t a smooth matrix function $Z(s, t)$ such that

$$\exp(sA)\exp(tB) = \exp Z(s, t). \tag{4.8}$$

Using Lemma 4.1, the derivative of (4.8) with respect to s is

$$A\,\exp(sA)\exp(tB) = d\exp_{Z(s,t)}\Big(\frac{\partial Z}{\partial s}(s, t)\Big)\exp Z(s, t),$$

so that

$$\frac{\partial Z}{\partial s} = d\exp_Z^{-1}(A) = A - \frac{1}{2}\,[Z, A] + \sum_{k\geq 2} \frac{B_k}{k!}\,\mathrm{ad}\,_Z^k(A). \tag{4.9}$$

We next take the inverse of (4.8)

---

[4] John Edward Campbell, born: 27 May 1862 in Lisburn, Co Antrim (Ireland), died: 1 October 1924 in Oxford (England).

[5] Henry Frederick Baker, born: 3 July 1866 in Cambridge (England), died: 17 March 1956 in Cambridge.

[6] Felix Hausdorff, born: 8 November 1869 in Breslau, Silesia (now Wroclaw, Poland), died: 26 January 1942 in Bonn (Germany).

$$\exp(-tB)\exp(-sA) = \exp\big(-Z(s,t)\big),$$

and differentiate this relation with respect to t. As above we get

$$\frac{\partial Z}{\partial t} = d\exp^{-1}_{-Z}(B) = B + \frac{1}{2}\,[Z,B] + \sum_{k\geq 2} \frac{B_k}{k!}\,\mathrm{ad}_Z^{\,k}(B), \tag{4.10}$$

because $\mathrm{ad}_{-Z}^{\,k}(B) = (-1)^k \mathrm{ad}_Z^{\,k}(B)$ and the Bernoulli numbers satisfy $B_k = 0$ for odd $k > 2$. A comparison of (4.6) with (4.8) gives $C(t) = Z(t,t)$. The stated differential equation for $C(t)$ therefore follows from $\dot C(t) = \frac{\partial Z}{\partial s}(t,t) + \frac{\partial Z}{\partial t}(t,t)$, and from adding the relations (4.9) and (4.10). □

Using Lemma 4.3 we can compute the first Taylor coefficients of $C(t)$,

$$\exp(tA)\exp(tB) = \exp\Big(tC_1 + t^2C_2 + t^3C_3 + t^4C_4 + t^5C_5 + \ldots\Big). \tag{4.11}$$

Inserting this expansion of $C(t)$ into (4.7) and comparing like powers of t gives

$$\begin{aligned}
C_1 &= A + B\\
C_2 &= \frac{1}{4}\,[A-B, A+B] = \frac{1}{2}\,[A,B]\\
C_3 &= \frac{1}{6}\Big[A-B, \frac{1}{2}\,[A,B]\Big] \;=\; \frac{1}{12}\Big[A,[A,B]\Big] + \frac{1}{12}\Big[B,[B,A]\Big]\\
C_4 &= \;\ldots\; = \frac{1}{24}\Big[A,\big[B,[B,A]\big]\Big]\\
C_5 &= \;\ldots\; = -\frac{1}{720}\Big[A,\big[A,\big[A,[A,B]\big]\big]\Big] - \frac{1}{720}\Big[B,\big[B,\big[B,[B,A]\big]\big]\Big]\\
&\quad + \frac{1}{360}\Big[A,\big[B,\big[B,[B,A]\big]\big]\Big] + \frac{1}{360}\Big[B,\big[A,\big[A,[A,B]\big]\big]\Big]\\
&\quad + \frac{1}{120}\Big[A,\big[A,\big[B,[B,A]\big]\big]\Big] + \frac{1}{120}\Big[B,\big[B,\big[A,[A,B]\big]\big]\Big].
\end{aligned} \tag{4.12}$$

Here, the dots $\ldots$ in the formulas for C_4 and C_5 indicate simplifications with the help of the Jacobi identity

$$\big[A,[B,C]\big] + \big[C,[A,B]\big] + \big[B,[C,A]\big] = 0, \tag{4.13}$$

which is verified by straightforward calculation. For higher order the expressions soon become very complicated.

The Symmetric BCH Formula. For the construction of symmetric splitting methods it is convenient to use a formula for the composition

$$\exp\Big(\frac{t}{2}A\Big)\exp(tB)\exp\Big(\frac{t}{2}A\Big) = \exp\Big(tS_1 + t^3S_3 + t^5S_5 + \ldots\Big). \tag{4.14}$$

Since the inverse of the left-hand side is obtained by changing the sign of t, the same must be true for the right-hand side. This explains why only odd powers of

t are present in (4.14). Applying the BCH formula (4.11) to $\exp(\frac{t}{2}A)\exp(\frac{t}{2}B) = \exp C(t)$ and a second time to $\exp(C(t))\exp(-C(-t))$ yields for the coefficients of (4.14) (Yoshida 1990)

$$\begin{aligned} S_1 &= A + B \\ S_3 &= -\frac{1}{24}\Big[A,[A,B]\Big] + \frac{1}{12}\Big[B,[B,A]\Big] \\ S_5 &= \frac{7}{5760}\Big[A,\big[A,\big[A,[A,B]\big]\big]\Big] - \frac{1}{720}\Big[B,\big[B,\big[B,[B,A]\big]\big]\Big] \\ &\quad + \frac{1}{360}\Big[A,\big[B,\big[B,[B,A]\big]\big]\Big] + \frac{1}{360}\Big[B,\big[A,\big[A,[A,B]\big]\big]\Big] \\ &\quad - \frac{1}{480}\Big[A,\big[A,\big[B,[B,A]\big]\big]\Big] + \frac{1}{120}\Big[B,\big[B,\big[A,[A,B]\big]\big]\Big]. \end{aligned} \tag{4.15}$$

III.5 Order Conditions via the BCH Formula

Using the BCH formula we present an alternative approach to the order conditions of splitting and composition methods. The main idea is to write the flow of a differential equation formally as the exponential of the Lie derivative.

III.5.1 Calculus of Lie Derivatives

For a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y),$$

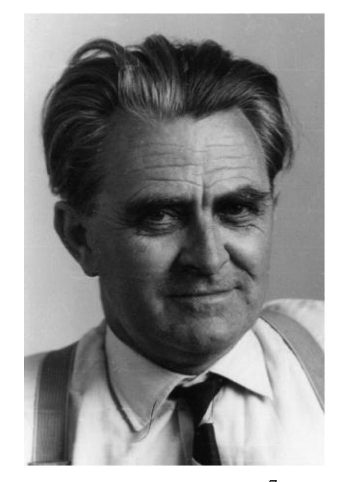
it is convenient to study the composition of the flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems

$$\dot{y} = f^{[1]}(y), \qquad \dot{y} = f^{[2]}(y), \tag{5.1}$$

respectively. We introduce the differential operators (*Lie derivative*)

$$D_i = \sum_j f_j^{[i]}(y)\,\frac{\partial}{\partial y_j}$$

which means that for differentiable functions $F : \mathbb{R}^n \to \mathbb{R}^m$ we have



Wolfgang Gröbner[7]

$$D_i F(y) = F'(y) f^{[i]}(y). \tag{5.2}$$

It follows from the chain rule that, for the solutions $\varphi_t^{[i]}(y_0)$ of (5.1),

[7] Wolfgang Gröbner, born: 11 February 1899 in Gossensass, South Tyrol (now Italy), died: 10 August 1980 in Innsbruck.

$$\frac{d}{dt} F\big(\varphi_t^{[i]}(y_0)\big) \;=\; \big(D_i F\big)\big(\varphi_t^{[i]}(y_0)\big), \tag{5.3}$$

and applying this operator iteratively we get

$$\frac{d^k}{dt^k} F\big(\varphi_t^{[i]}(y_0)\big) \;=\; \big(D_i^k F\big)\big(\varphi_t^{[i]}(y_0)\big). \tag{5.4}$$

Consequently, the Taylor series of $F\big(\varphi_t^{[i]}(y_0)\big)$, developed at $t=0$, becomes

$$F\big(\varphi_t^{[i]}(y_0)\big) \;=\; \sum_{k\ge 0} \frac{t^k}{k!}\,(D_i^k F)(y_0) \;=\; \exp(tD_i)F(y_0). \tag{5.5}$$

Now, putting $F(y) = \mathrm{Id}(y) = y$, the identity map, this is the Taylor series of the solution itself

$$\varphi_t^{[i]}(y_0) \;=\; \sum_{k\ge 0} \frac{t^k}{k!}\,(D_i^k \mathrm{Id})(y_0) \;=\; \exp(tD_i)\mathrm{Id}(y_0). \tag{5.6}$$

If the functions $f^{[i]}(y)$ are not analytic, but only N-times continuously differentiable, the series (5.6) has to be truncated and a $\mathcal{O}(h^N)$ remainder term has to be included.

Lemma 5.1 (Gröbner 1960). *Let $\varphi_s^{[1]}$ and $\varphi_t^{[2]}$ be the flows of the differential equations $\dot y = f^{[1]}(y)$ and $\dot y = f^{[2]}(y)$, respectively. For their composition we then have*

$$\Big(\varphi_t^{[2]} \circ \varphi_s^{[1]}\Big)(y_0) = \exp(sD_1)\;\exp(tD_2)\,\mathrm{Id}(y_0).$$

Proof. This is precisely formula (5.5) with $i=1$, t replaced with s, and with $F(y) = \varphi_t^{[2]}(y) = \exp(tD_2)\mathrm{Id}(y_0)$. □

Remark 5.2. Notice that the indices 1 and 2 as well as s and t to the left and right in the identity of Lemma 5.1 are permuted. Gröbner calls this phenomenon, which sometimes leads to some confusion in the literature, the "Vertauschungssatz".

Remark 5.3. The statement of Lemma 5.1 can be extended to more than two flows. If $\varphi_t^{[j]}$ is the flow of a differential equation $\dot y = f^{[j]}(y)$, then we have

$$\Big(\varphi_u^{[m]} \circ \ldots \circ \varphi_t^{[2]} \circ \varphi_s^{[1]}\Big)(y_0) = \exp(sD_1)\;\exp(tD_2)\cdot\ldots\cdot\exp(uD_m)\mathrm{Id}(y_0).$$

This follows by induction on m.

In general, the two operators D_1 and D_2 do not commute, so that the composition $\exp(tD_1)\exp(tD_2)\mathrm{Id}(y_0)$ is different from $\exp\big(t(D_1+D_2)\big)\mathrm{Id}(y_0)$, which represents the solution $\varphi_t(y_0)$ of $\dot y = f(y) = f^{[1]}(y) + f^{[2]}(y)$. The relation of Lemma 5.1 suggests the use of the BCH formula. However, D_1 and D_2 are unbounded differential operators so that the series expansions that appear cannot be

expected to converge. A formal application of the BCH formula with tA and tB replaced with sD_1 and tD_2, respectively, yields

$$\exp(sD_1)\exp(tD_2) = \exp\big(D(s,t)\big), \tag{5.7}$$

where the differential operator $D(s,t)$ is obtained from (4.11) as

$$\begin{aligned} D(s,t) \;&=\; sD_1 + tD_2 + \frac{st}{2}[D_1,D_2] + \frac{s^2t}{12}\Big[D_1,[D_1,D_2]\Big] \\ &\quad + \frac{st^2}{12}\Big[D_2,[D_2,D_1]\Big] + \frac{s^2t^2}{24}\Big[D_1,\big[D_2,[D_2,D_1]\big]\Big] + \ldots\,. \end{aligned} \tag{5.8}$$

The *Lie bracket* for differential operators is calculated exactly as for matrices, namely, $[D_1,D_2] = D_1D_2 - D_2D_1$. But how can we interpret (5.7) rigorously? Expanding both sides in Taylor series we see that

$$\exp(sD_1)\exp(tD_2) = I + sD_1 + tD_2 + \frac{1}{2}\Big(s^2D_1^2 + 2stD_1D_2 + t^2D_2^2\Big) + \ldots \tag{5.9}$$

and

$$\begin{aligned} \exp\big(D(s,t)\big) \;&=\; I + D(s,t) + \frac{1}{2}D(s,t)^2 + \ldots \\ &=\; I + sD_1 + tD_2 + \frac{1}{2}\Big((sD_1+tD_2)^2 + st[D_1,D_2]\Big) + \ldots\,. \end{aligned}$$

By derivation of the BCH formula we have a formal identity, i.e., both series have exactly the same coefficients. Moreover, every finite truncation of the series can be applied without any difficulties to sufficiently differentiable functions $F(y)$. Consequently, for N-times differentiable functions the relation (5.7) holds true, if both sides are replaced by their truncated Taylor series and if a $\mathcal{O}(h^N)$ remainder is added ($h = \max(|s|,|t|)$).

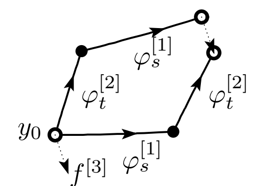
III.5.2 Lie Brackets and Commutativity

If we apply D_2 to a function F, followed by an application of D_1, we will obtain partial derivatives of F of first and second orders. However, if we subtract from this the same expression with D_1 and D_2 reversed, the second derivatives will cancel (this was already remarked upon by Jacobi (1862), p. 39: "differentialia partialia secunda functionis f non continere") and we see that the Lie bracket

$$[D_1,D_2] = D_1D_2 - D_2D_1 = \sum_i \Big(\sum_j \Big(\frac{\partial f_i^{[2]}}{\partial y_j}\, f_j^{[1]} \;-\; \frac{\partial f_i^{[1]}}{\partial y_j}\, f_j^{[2]}\Big)\Big)\frac{\partial}{\partial y_i} \tag{5.10}$$

is again a linear differential operator. So, from two vector fields $f^{[1]}$ and $f^{[2]}$ we obtain a *third* vector field $f^{[3]}$.

The *geometric meaning* of the new vector field can be deduced from Lemma 5.1. We see by subtracting (5.9) from itself, once as it stands and once with sD_1 and tD_2 permuted, that



$$\varphi_t^{[2]}\circ\varphi_s^{[1]}(y_0)-\varphi_s^{[1]}\circ\varphi_t^{[2]}(y_0) = st\,[D_1,D_2]\,\mathrm{Id}(y_0)+\ldots = st\,f^{[3]}(y_0)+\ldots \quad (5.11)$$

(see the picture), where "$+\ldots$" are terms of order ≥ 3. This leads us to the following result.

Lemma 5.4. *Let $f^{[1]}(y)$ and $f^{[2]}(y)$ be defined on an open set. The corresponding flows $\varphi_s^{[1]}$ and $\varphi_t^{[2]}$ commute everywhere for all sufficiently small s and t, if and only if*

$$[D_1, D_2] = 0. \quad (5.12)$$

Proof. The "only if" part is clear from (5.11). For proving the "if" part, we take s and t fixed, and subdivide, for a given n, the integration intervals into n equidistant parts $\Delta s = s/n$ and $\Delta t = t/n$. This allows us to transform the solution $\varphi_t^{[2]}\circ\varphi_s^{[1]}(y_0)$ by a discrete homotopy in n^2 steps into the solution $\varphi_s^{[1]}\circ\varphi_t^{[2]}(y_0)$, each time appending a small rectangle of size $\mathcal{O}(n^{-2})$. If we denote such an intermediate stage by

$$\Gamma_k = \ldots\circ\varphi^{[2]}_{j_2\Delta t}\circ\varphi^{[1]}_{i_2\Delta s}\circ\varphi^{[2]}_{j_1\Delta t}\circ\varphi^{[1]}_{i_1\Delta s}(y_0)$$

then we have $\Gamma_0 = \varphi_t^{[2]}\circ\varphi_s^{[1]}(y_0)$ and $\Gamma_{n^2} = \varphi_s^{[1]}\circ\varphi_t^{[2]}(y_0)$ (see Fig. 5.1). Now, for $n\to\infty$, we have the estimate

$$|\Gamma_{k+1}-\Gamma_k| \leq \mathcal{O}(n^{-3}),$$

because the error terms in (5.11) are of order 3 at least, and because of the differentiability of the solutions with respect to initial values. Thus, by the triangle inequality $|\Gamma_{n^2}-\Gamma_0| \leq \mathcal{O}(n^{-1})$ and the result is proved. □

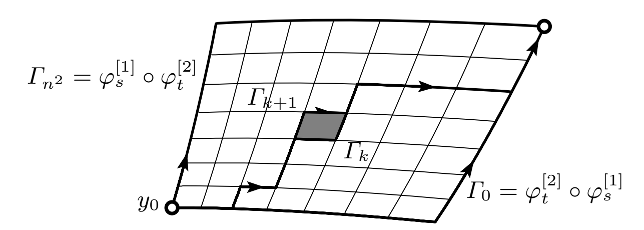


Fig. 5.1. Estimation of commuting solutions

III.5.3 Splitting Methods

We follow the approach of Yoshida (1990) for obtaining the order conditions of splitting methods (II.5.6). The idea is the following: with the use of Lemma 5.1 we write the method as a product of exponentials, then we apply formally the Baker-Campbell-Hausdorff formula to get one exponential of a series in powers of h. Finally, we compare this series with $h(D_1 + D_2)$, which corresponds to the exact solution of (5.1).

The splitting method (II.5.6), viz.,

$$\Psi_h = \varphi^{[2]}_{b_m h} \circ \varphi^{[1]}_{a_m h} \circ \varphi^{[2]}_{b_{m-1} h} \circ \ldots \circ \varphi^{[1]}_{a_2 h} \circ \varphi^{[2]}_{b_1 h} \circ \varphi^{[1]}_{a_1 h}, \tag{5.13}$$

is a composition of expressions $\varphi^{[2]}_{b_j h} \circ \varphi^{[1]}_{a_j h}$ which, by Lemma 5.1 and by (5.7), can be written as an exponential

$$\begin{aligned}\varphi^{[2]}_{b_j h} \circ \varphi^{[1]}_{a_j h} = \exp\Big(&a_j h E_1^1 + b_j h E_2^1 + a_j b_j h^2 E_1^2 \\ &+ a_j^2 b_j h^3 E_1^3 + a_j b_j^2 h^3 E_2^3 + a_j^2 b_j^2 h^4 E_1^4 + \ldots\Big)\mathrm{Id},\end{aligned} \tag{5.14}$$

where we use the abbreviations

$$E_1^1 = D_1, \quad E_2^1 = D_2, \quad E_1^2 = \frac{1}{2}[D_1, D_2], \quad E_1^3 = \frac{1}{12}\big[D_1, [D_1, D_2]\big],$$
$$E_2^3 = \frac{1}{12}\big[D_2, [D_2, D_1]\big], \quad E_1^4 = \frac{1}{24}\big[D_1[D_2, [D_2, D_1]]\big],$$

and the dots indicate $\mathcal{O}(h^5)$ expressions.

We next define $\Psi^{(j)}$ recursively by

$$\Psi^{(0)} = \mathrm{Id}, \qquad \Psi^{(j)} = \varphi^{[2]}_{b_j h} \circ \varphi^{[1]}_{a_j h} \circ \Psi^{(j-1)}, \tag{5.15}$$

so that $\Psi^{(m)}$ is equal to our method (5.13). Aiming to write $\Psi^{(j)}$ also as an exponential of differential operators, we are confronted with computing commutators of the expressions E_i^j. We see that $[E_1^1, E_2^1] = 2E_1^2$, $[E_1^1, E_1^2] = 6E_1^3$, $[E_2^1, E_1^2] = -6E_2^3$, $[E_1^1, E_2^3] = 2E_1^4$, and $[E_2^1, \alpha_1^3] = -2E_1^4$ as a consequence of the Jacobi identity (4.13). But the other commutators cannot be expressed in terms of E_i^j. We therefore introduce

$$E_2^4 = \frac{1}{24}\big[D_1, [D_1, [D_1, D_2]]\big], \quad E_3^4 = \frac{1}{24}\big[D_2, [D_2, [D_2, D_1]]\big].$$

This allows us to formulate the following result.

Lemma 5.5. *The method $\Psi^{(j)}$, defined by (5.15), can be formally written as*

$$\begin{aligned}\Psi^{(j)} = \exp\Big(&c_{1,j}^1 h E_1^1 + c_{2,j}^1 h E_2^1 + c_{1,j}^2 h^2 E_1^2 + c_{1,j}^3 h^3 E_1^3 \\ &+ c_{2,j}^3 h^3 E_2^3 + c_{1,j}^4 h^4 E_1^4 + c_{2,j}^4 h^4 E_2^4 + c_{3,j}^4 h^4 E_3^4 + \ldots\Big)\mathrm{Id},\end{aligned}$$

where all coefficients are zero for $j = 0$, and where for $j \geq 1$

$$
\begin{aligned}
c_{1,j}^1 &= c_{1,j-1}^1 + a_j, \qquad c_{2,j}^1 = c_{2,j-1}^1 + b_j, \\
c_{1,j}^2 &= c_{1,j-1}^2 + a_j b_j + c_{1,j-1}^1 b_j - c_{2,j-1}^1 a_j, \\
c_{1,j}^3 &= c_{1,j-1}^3 + a_j^2 b_j + 2c_{1,j-1}^1 a_j b_j - 3c_{1,j-1}^2 a_j \\
&\qquad + (c_{1,j-1}^1)^2 b_j - c_{1,j-1}^1 c_{2,j-1}^1 a_j + c_{2,j-1}^1 a_j^2, \\
c_{2,j}^3 &= c_{2,j-1}^3 + a_j b_j^2 - 4c_{2,j-1}^1 a_j b_j + 3c_{1,j-1}^2 b_j \\
&\qquad + (c_{2,j-1}^1)^2 a_j - c_{1,j-1}^1 c_{2,j-1}^1 b_j + c_{1,j-1}^1 b_j^2,
\end{aligned}
$$

and similar but more complicated formulas for $c_{i,j}^4$.

Proof. Due to the reversed order in Lemma 5.1 we have to compute $\exp(A)\exp(B)$, where A is the argument of the exponential for $\Psi^{(j-1)}$ and B is that of (5.14). The rest is a tedious but straightforward application of the BCH formula. One has to use repeatedly the formulas for $[E_i^j, E_k^l]$, stated before Lemma 5.5. □

Theorem 5.6. *The splitting method (5.13) is of order p if*

$$
c_{1,m}^1 = c_{2,m}^1 = 1\,, \qquad c_{\ell,m}^k = 0 \quad \textit{for} \;\; k = 2, \ldots, p \;\; \textit{and all} \;\; \ell. \tag{5.16}
$$

The coefficients $c_{\ell,m}^k$ are those defined in Lemma 5.5.

Proof. This is an immediate consequence of Lemma 5.5, because the conditions of order p imply that the Taylor series expansion of $\Psi^{(m)}(y_0)$ coincides with that of the solution $\varphi_h(y_0) = \exp\big(h(D_1 + D_2)\big)y_0$ up to terms of size $\mathcal{O}(h^p)$. □

A simplification in the order conditions arises for symmetric methods (5.13), that is, for coefficients satisfying $a_{m+1-i} = a_i$ and $b_{m-i} = b_i$ for all i (and $b_m = 0$). By Theorem II.3.2, it is sufficient to consider the order conditions (5.16) for odd k only.

III.5.4 Composition Methods

We now consider composition methods (II.4.6), viz.,

$$
\Psi_h = \Phi_{\alpha_s h} \circ \Phi^*_{\beta_s h} \circ \ldots \circ \Phi^*_{\beta_2 h} \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}, \tag{5.17}
$$

where Φ_h is a first-order method for $\dot y = f(y)$ and Φ_h^* is its adjoint. We assume

$$
\Phi_h = \exp\Big(hC_1 + h^2 C_2 + h^3 C_3 + \ldots\Big)\mathrm{Id} \tag{5.18}
$$

with differential operators C_i, and such that C_1 is the Lie derivative operator corresponding to $\dot y = f(y)$. For the splitting method $\Phi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$ this follows from (5.14), and for general one-step methods this is a consequence of Sect. IX.1 on backward error analysis. The adjoint method then satisfies

$$\Phi_h^* = \exp\Big(hC_1 - h^2C_2 + h^3C_3 - \ldots\Big)\mathrm{Id}. \tag{5.19}$$

From now on the procedure is similar to that of Sect. III.5.3. We define $\Psi^{(j)}$ recursively by

$$\Psi^{(0)} = \mathrm{Id}, \qquad \Psi^{(j)} = \Phi_{\alpha_j h} \circ \Phi^*_{\beta_j h} \circ \Psi^{(j-1)}, \tag{5.20}$$

so that $\Psi^{(m)}$ becomes (5.17). We apply the BCH formula to obtain

$$\begin{aligned}\Phi_{\alpha_j h} \circ \Phi^*_{\beta_j h} &= \exp\Big(\beta_j hC_1 - \beta_j^2h^2C_2 + \ldots\Big)\exp\Big(\alpha_j hC_1 + \alpha_j^2h^2C_2 + \ldots\Big)\mathrm{Id} \\ &= \exp\Big((\alpha_j+\beta_j)hE_1^1 + (\alpha_j^2-\beta_j^2)h^2E_1^2 \\ &\qquad + (\alpha_j^3+\beta_j^3)h^3E_1^3 + \frac{1}{2}\alpha_j\beta_j(\alpha_j+\beta_j)h^3E_2^3 + \ldots\Big)\mathrm{Id}\end{aligned}$$

where

$$E_1^k = C_k, \qquad E_2^3 = [C_1, C_2].$$

We then have the following result.

Lemma 5.7. *The method $\Psi^{(j)}$ of (5.20) can be formally written as*

$$\Psi^{(j)} = \exp\Big(\gamma_{1,j}^1 hE_1^1 + \gamma_{1,j}^2 h^2E_1^2 + \gamma_{1,j}^3 h^3E_1^3 + \gamma_{2,j}^3 h^3E_2^3 + \ldots\Big)\mathrm{Id},$$

where all coefficients are zero for $j = 0$, and where for $j = 1, \ldots, m$

$$\begin{aligned}\gamma_{1,j}^1 &= \gamma_{1,j-1}^1 + \alpha_j + \beta_j \\ \gamma_{1,j}^2 &= \gamma_{1,j-1}^2 + \alpha_j^2 - \beta_j^2 \\ \gamma_{1,j}^3 &= \gamma_{1,j-1}^3 + \alpha_j^3 + \beta_j^3 \\ \gamma_{2,j}^3 &= \gamma_{2,j-1}^3 + \frac{1}{2}\alpha_j\beta_j(\alpha_j+\beta_j) + \frac{1}{2}\gamma_{1,j-1}^1(\alpha_j^2-\beta_j^2) - \frac{1}{2}\gamma_{1,j-1}^2(\alpha_j+\beta_j).\end{aligned}$$

Proof. Similar to Lemma 5.5, the result follows using the BCH formula. □

Theorem 5.8. *The composition method (5.17) is of order p if*

$$\gamma_{1,m}^1 = 1\,, \qquad \gamma_{\ell,m}^k = 0 \quad \textit{for } k = 2, \ldots, p \textit{ and all } \ell. \tag{5.21}$$

The coefficients $\gamma_{\ell,m}^k$ are those defined in Lemma 5.7. □

It is interesting to see how these order conditions are related to those obtained with the use of trees. The conditions $\gamma_{1,m}^1 = 1$ and $\gamma_{1,m}^2 = \gamma_{1,m}^3 = 0$ are identical to the first three order conditions of Example 3.15. The remaining condition for order 3, $\gamma_{2,m}^3 = 0$, reads

$$\begin{aligned}&\sum_{k=1}^m \alpha_k\beta_k(\alpha_k+\beta_k) + \sum_{k=1}^m(\alpha_k^2-\beta_k^2)\sum_{i=1}^{k-1}(\alpha_i+\beta_i) - \sum_{k=1}^m(\alpha_k+\beta_k)\sum_{i=1}^{k-1}(\alpha_i^2-\beta_i^2) \\ &\qquad = \sum_{k=1}^m(\alpha_k^2-\beta_k^2)\sum_{i=1}^{k}{}'(\alpha_i+\beta_i) - \sum_{k=1}^m(\alpha_k+\beta_k)\sum_{i=1}^{k}{}'(\alpha_i^2-\beta_i^2) = 0.\end{aligned}$$

This condition is just the difference of the order conditions for the trees ② ∘ ① and ① ∘ ②, whose sum is zero by the Switching Lemma 3.8. Therefore the condition $\gamma_{2,m}^3 = 0$ is equivalent to (though more complicated than) the fourth condition of Example 3.15.

Symmetric Composition of Symmetric Methods. Consider now a composition

$$\Psi_h = \Phi_{\gamma_m h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h} \circ \Phi_{\gamma_2 h} \circ \ldots \circ \Phi_{\gamma_m h}, \tag{5.22}$$

where Φ_h is a symmetric method that can be written as

$$\Phi_h = \exp\Big(hS_1 + h^3 S_3 + h^5 S_5 + \ldots\Big)\mathrm{Id}$$

with S_1 the Lie derivative operator corresponding to $\dot{y} = f(y)$. For the Strang splitting $\Phi_h = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ such an expansion follows from the symmetric BCH formula (4.14), and for general symmetric one-step methods from Sect. IX.2. The derivation of the order conditions is similar to the above with $\Psi^{(j)}$ defined by

$$\Psi^{(1)} = \Phi_{\gamma_1 h}, \qquad \Psi^{(j)} = \Phi_{\gamma_j h} \circ \Psi^{(j-1)} \circ \Phi_{\gamma_j h},$$

so that $\Psi^{(m)}$ becomes (5.22).

Lemma 5.9. *The method $\Psi^{(j)}$ can be formally written as*

$$\Psi^{(j)} = \exp\Big(\sigma_{1,j}^1 h E_1^1 + \sigma_{1,j}^3 h^3 E_1^3 + \sigma_{1,j}^5 h^5 E_1^5 + \sigma_{2,j}^5 h^5 E_2^5 + \ldots\Big)\mathrm{Id},$$

where $E_1^k = S_k$, $E_2^5 = \big[S_1[S_1, S_3]\big]$, and where $\sigma_{1,1}^k = \gamma_1^k$, $\sigma_{2,1}^5 = 0$, and

$$\begin{aligned}
\sigma_{1,j}^k &= \sigma_{1,j-1}^k + 2\gamma_j^k \\
\sigma_{2,j}^5 &= \sigma_{2,j-1}^5 + \frac{1}{6}\Big(\gamma_j^3 (\sigma_{1,j-1}^1)^2 - \gamma_j \sigma_{1,j-1}^1 \sigma_{1,j-1}^3 - \gamma_j^2 \sigma_{1,j-1}^3 + \gamma_j^4 \sigma_{1,j-1}^1\Big).
\end{aligned}$$

Proof. The result is a consequence of the symmetric BCH formula (4.14) with $\gamma_j h S_1 + \gamma_j^3 h^3 S_3 + \ldots$ and $\sigma_{1,j-1}^1 h E_1^1 + \sigma_{1,j-1}^3 h E_1^3 + \ldots$ in the roles of $\frac{t}{2}A$ and tB, respectively. □

Theorem 5.10. *The composition method (5.22) is of order p if*

$$\sigma_{1,m}^1 = 1\,, \qquad \sigma_{\ell,m}^k = 0 \quad \textit{for odd} \;\; k = 3, \ldots, p \;\; \textit{and all} \;\; \ell. \tag{5.23}$$

The coefficients $\sigma_{\ell,m}^k$ are those defined in Lemma 5.9. □

Symmetric composition methods up to order 10 will be constructed and discussed in Sect. V.3.

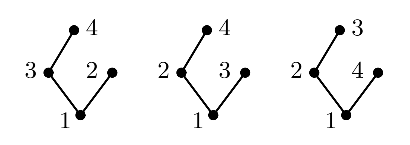
III.6 Exercises

1. Find all trees of orders 5 and 6.
2. (A. Cayley 1857). Denote the number of trees of order q by a_q. Prove that

$$a_1 + a_2 x + a_3 x^2 + a_4 x^3 + \ldots = (1-x)^{-a_1}(1-x^2)^{-a_2}(1-x^3)^{-a_3} \cdot \ldots .$$

q	1	2	3	4	5	6	7	8	9	10
a_q	1	1	2	4	9	20	48	115	286	719

3. Independency of the elementary differentials: show that for every $\tau \in T$ there is a system (1.1) such that the first component of $F(\tau)(0)$ equals 1, and the first component of $F(u)(0)$ is zero for all trees $u \neq t$.
Hint. Consider a monotonic labelling of τ, and define y_i' as the product over all y_j, where j runs through all labels of vertices that lie directly above the vertex "i". For the first labelling of the tree of Exercise 4 this would be $\dot{y}_1 = y_2 y_3$, $\dot{y}_2 = 1$, $\dot{y}_3 = y_4$, and $\dot{y}_4 = 1$.
4. Prove that the coefficient $\alpha(\tau)$ of Definition 1.2 is equal to the number of possible monotonic labellings of the vertices of τ, starting with the label 1 for the root. For example, the tree $[[\bullet], \bullet]$ has three different monotonic labellings.



In addition, deduce, from (1.22), the recursion formula

$$\alpha(\tau) = \binom{|\tau|-1}{|\tau_1|, \ldots, |\tau_m|} \alpha(\tau_1) \cdot \ldots \cdot \alpha(\tau_m) \frac{1}{\mu_1! \mu_2! \ldots}, \tag{6.1}$$

where the integers $\mu_1, \mu_2, \ldots$ count equal trees among $\tau_1, \ldots, \tau_m$ and

$$\binom{|\tau|-1}{|\tau_1|, \ldots, |\tau_m|} = \frac{(|\tau|-1)!}{|\tau_1|! \cdot \ldots \cdot |\tau_m|!}$$

denotes the multinomial coefficient.
Remark. In the theoretical physics literature, the coefficients $\alpha(\tau)$ are written *CM*(τ) and called "Connes-Moscovici weights".
5. If we denote by $N(\tau)$ the number of elements in *OST*(τ), then show that

$$N(\bullet) = 2, \qquad N([\tau_1, \ldots, \tau_m]) = 1 + N(\tau_1) \cdot \ldots \cdot N(\tau_m).$$

Use this result to compute the number of subtrees of the christmas tree decorating formula (1.34). *Answer:* 6865.
6. Prove that the elementary differentials for partitioned problems are independent. For a given tree ($\tau \in TP$, find a problem (2.1) such that a certain component of $F(\tau)(p, q)$ vanishes for all $u \in TP$ except for τ.
Hint. Consider the construction of Exercise 3, and define the partitioning of y into (p, q) according to the colours of the vertices.

7. The number of order conditions for partitioned Runge–Kutta methods (II.2.2) is $2a_r$ for order r, where a_r is given by (see Hairer, Nørsett & Wanner (1993), page 311)

r	1	2	3	4	5	6	7	8	9	10
a_r	1	2	7	26	107	458	2058	9498	44987	216598

Find a formula similar to that of Exercise 2.

8. For the special second order differential equation $\ddot{y} = g(y)$, and for a Nyström method

$$\begin{aligned} \ell_i &= g\Big(y_0 + c_i h\, \dot{y}_0 + h^2 \sum_{j=1}^s a_{ij}\ell_j\Big), \\ y_1 &= y_0 + h\dot{y}_0 + h^2 \sum_{i=1}^s \beta_i \ell_i, \qquad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^s b_i \ell_i, \end{aligned} \tag{6.2}$$

consider the simplifying assumption

$$\begin{aligned} CN(\eta): &\quad \sum_{j=1}^s a_{ij} c_j^{k-2} = \frac{c_i^k}{k(k-1)}, \quad k = 2, \ldots, \eta, \\ DN(\zeta): &\quad \sum_{i=1}^s b_i c_i^{k-2} a_{ij} = b_j \Big(\frac{c_j^k}{k(k-1)} - \frac{c_j}{k-1} + \frac{1}{k}\Big), \quad k = 2, \ldots, \zeta. \end{aligned}$$

Prove that if the quadrature formula (b_i, c_i) is of order p, if $\beta_i = b_i(1 - c_i)$ for all i, and if the simplifying assumptions $CN(\eta)$, $DN(\zeta)$ are satisfied with $2\eta + 2 \ge p$ and $\zeta + \eta \ge p$, then the Nyström method has order p.

9. *Nyström methods of maximal order* $2s$. Prove that there exists a one-parameter family of s-stage Nyström methods (6.2) for $\ddot{y} = g(y)$, which have order $2s$. *Hint.* Consider the Gaussian quadrature formula and define the coefficients a_{ij} by $CN(s)$ and by

$$\sum_{i=1}^s b_i c_i^{k-2} a_{is} = b_j \Big(\frac{c_s^k}{k(k-1)} - \frac{c_s}{k-1} + \frac{1}{k}\Big)$$

for $k = 2, \ldots, s$.

10. Prove that the coefficient C_4 in the series (4.11) of the Baker-Campbell-Hausdorff formula is given by $C_4 = [A, [B, [B, A]]]/24$.

11. Prove that the series (4.11) converges for $|t| < \ln 2/(\|A\| + \|B\|)$.

12. By Theorem 5.10 four order conditions have to be satisfied such that the symmetric composition method (5.22) is of order 6. Prove that these conditions are equivalent to the four conditions of Example V.3.15. (Care has to be taken due to the different meaning of the γ_i.)

Chapter IV.
Conservation of First Integrals and Methods on Manifolds

This chapter deals with the conservation of invariants (first integrals) by numerical methods, and with numerical methods for differential equations on manifolds. Our investigation will follow two directions. We first investigate which of the methods introduced in Chap. II conserve invariants automatically. We shall see that most of them conserve linear invariants, a few of them quadratic invariants, and none of them conserves cubic or general nonlinear invariants. We then construct new classes of methods, which are adapted to known invariants and which force the numerical solution to satisfy them. In particular, we study projection methods and methods based on local coordinates of the manifold defined by the invariants. We discuss in some detail the case where the manifold is a Lie group. Finally, we consider differential equations on manifolds with orthogonality constraints, which often arise in numerical linear algebra.

IV.1 Examples of First Integrals

> Je nomme intégrale une équation $u = Const.$ telle que sa différentielle $du = 0$ soit vérifiée identiquement par le système des équations différentielles proposées . . . (C.G.J. Jacobi 1840, p. 350)

We consider differential equations

$$\dot{y} = f(y), \tag{1.1}$$

where y is a vector or possibly a matrix.

Definition 1.1. A non-constant function $I(y)$ is called a *first integral* of (1.1) if

$$I'(y)f(y) = 0 \qquad \text{for all } y. \tag{1.2}$$

This implies that *every* solution $y(t)$ of (1.1) satisfies $I\big(y(t)\big) = I(y_0) = Const.$ Synonymously with “first integral”, the terms *invariant* or *conserved quantity* or *constant of motion* are also used.

In Chap. I we have seen many examples of differential equations with invariants. For example, the Lotka–Volterra problem (I.1.1) has $I(u, v) = \ln u - u + 2 \ln v - v$ as first integral. The pendulum equation (I.1.13) has $H(p, q) = p^2/2 - \cos q$, and the Kepler problem (I.2.2) has two first integrals, namely H and L of (I.2.3) and (I.2.4).

Example 1.2 (Conservation of the Total Energy). Hamiltonian systems are of the form

$$\dot{p} = -H_q(p,q), \qquad \dot{q} = H_p(p,q),$$

where $H_q = \nabla_q H = \big(\partial H/\partial q\big)^T$ and $H_p = \nabla_p H = \big(\partial H/\partial p\big)^T$ are the column vectors of partial derivatives. The Hamiltonian function $H(p,q)$ is a first integral. This follows at once from $H'(p,q) = \big(\partial H/\partial p, \partial H/\partial q\big)$ and

$$\frac{\partial H}{\partial p}\Big(-\frac{\partial H}{\partial q}\Big)^T + \frac{\partial H}{\partial q}\Big(\frac{\partial H}{\partial p}\Big)^T = 0.$$

Example 1.3 (Conservation of the Total Linear and Angular Momentum of N-Body Systems). We consider a system of N particles interacting pairwise with potential forces which depend on the distances of the particles. This is formulated as a Hamiltonian system with total energy (I.4.1), viz.,

$$H(p,q) = \frac{1}{2}\sum_{i=1}^{N} \frac{1}{m_i}\, p_i^T p_i + \sum_{i=2}^{N}\sum_{j=1}^{i-1} V_{ij}\Big(\|q_i - q_j\|\Big)\,.$$

Here $q_i, p_i \in \mathbb{R}^3$ represent the position and momentum of the ith particle of mass m_i, and $V_{ij}(r)$ $(i > j)$ is the interaction potential between the ith and jth particle. The equations of motion read

$$\dot{q}_i = \frac{1}{m_i}\, p_i\,, \qquad \dot{p}_i = \sum_{j=1}^{N} \nu_{ij}\,(q_i - q_j)$$

where, for $i > j$, we have $\nu_{ij} = \nu_{ji} = -V'_{ij}(r_{ij})/r_{ij}$ with $r_{ij} = \|q_i - q_j\|$, and ν_{ii} is arbitrary, say $\nu_{ii} = 0$. The conservation of the total *linear momentum* $P = \sum_{i=1}^N p_i$ and the *angular momentum* $L = \sum_{i=1}^N q_i \times p_i$ is a consequence of the symmetry relation $\nu_{ij} = \nu_{ji}$:

$$\frac{d}{dt}\sum_{i=1}^{N} p_i = \sum_{i=1}^{N}\sum_{j=1}^{N} \nu_{ij}(q_i - q_j) = 0$$

$$\frac{d}{dt}\sum_{i=1}^{N} q_i \times p_i = \sum_{i=1}^{N} \frac{1}{m_i}\, p_i \times p_i + \sum_{i=1}^{N}\sum_{j=1}^{N} q_i \times \nu_{ij}(q_i - q_j) = 0\,.$$

Example 1.4 (Conservation of Mass in Chemical Reactions). Suppose that three substances A, B, C undergo a chemical reaction such as[1]

$$\begin{array}{lcll} A & \xrightarrow{0.04} & B & \text{(slow)} \\ B+B & \xrightarrow{3\cdot 10^7} & C+B & \text{(very fast)} \\ B+C & \xrightarrow{10^4} & A+C & \text{(fast).} \end{array}$$

[1] This *Robertson problem* is very popular in testing codes for stiff differential equations.

We denote the masses (or concentrations) of the substances A, B, C by y_1, y_2, y_3, respectively. By the mass action law this leads to the equations

$$\begin{array}{lrcl} \text{A:} & \dot y_1 &=& -0.04\,y_1 + 10^4\,y_2y_3 \\ \text{B:} & \dot y_2 &=& 0.04\,y_1 - 10^4\,y_2y_3 - 3\cdot 10^7\,y_2^2 \\ \text{C:} & \dot y_3 &=& 3\cdot 10^7\,y_2^2 \end{array}$$

We see that $\dot y_1 + \dot y_2 + \dot y_3 = 0$, hence the total mass $I(y) = y_1 + y_2 + y_3$ is an invariant of the system.

As was noted by Shampine (1986), such linear invariants are generally conserved by numerical integrators.

Theorem 1.5 (Conservation of Linear Invariants). *All explicit and implicit Runge–Kutta methods conserve linear invariants. Partitioned Runge–Kutta methods (II.2.2) conserve linear invariants if* $b_i = \widehat b_i$ *for all* i*, or if the invariant depends only on* p *or only on* q*.*

Proof. Let $I(y) = d^T y$ with a constant vector d, so that $d^T f(y) = 0$ for all y. In the case of Runge–Kutta methods we thus have $d^T k_i = 0$, and consequently $d^T y_1 = d^T y_0 + h d^T(\sum_{i=1}^s b_i k_i) = d^T y_0$. The statement for partitioned methods is proved similarly. □

Next we consider differential equations of the form

$$\dot Y = A(Y)Y, \tag{1.3}$$

where Y can be a vector or a matrix (not necessarily a square matrix). We then have the following result.

Theorem 1.6. *If* $A(Y)$ *is skew-symmetric for all* Y *(i.e.,* $A^T = -A$*), then the quadratic function* $I(Y) = Y^T Y$ *is an invariant. In particular, if the initial value* Y_0 *consists of orthonormal columns (i.e.,* $Y_0^T Y_0 = I$*), then the columns of the solution* $Y(t)$ *of (1.3) remain orthonormal for all* t*.*

Proof. The derivative of $I(Y)$ is $I'(Y)H = Y^T H + H^T Y$. Thus, we have $I'(Y)f(Y) = I'(Y)(A(Y)Y) = Y^T A(Y)Y + Y^T A(Y)^T Y$ for all Y which vanishes, because $A(Y)$ is skew-symmetric. This proves the statement. □

Example 1.7 (Rigid Body). The motion of a free rigid body, whose centre of mass is at the origin, is described by the Euler equations

$$\begin{array}{rclrcl} \dot y_1 &=& a_1 y_2 y_3, \qquad & a_1 &=& (I_2 - I_3)/(I_2 I_3) \\ \dot y_2 &=& a_2 y_3 y_1, \qquad & a_2 &=& (I_3 - I_1)/(I_3 I_1) \\ \dot y_3 &=& a_3 y_1 y_2, \qquad & a_3 &=& (I_1 - I_2)/(I_1 I_2) \end{array} \tag{1.4}$$

where the vector $y = (y_1, y_2, y_3)^T$ represents the angular momentum in the body frame, and I_1, I_2, I_3 are the principal moments of inertia (Euler (1758b); see Sect. VII.5 for a detailed description. This problem can be written as

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & y_3/I_3 & -y_2/I_2 \\ -y_3/I_3 & 0 & y_1/I_1 \\ y_2/I_2 & -y_1/I_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \tag{1.5}$$

which is of the form (1.3) with a skew-symmetric matrix $A(Y)$. By Theorem 1.6, $y_1^2 + y_2^2 + y_3^2$ is an invariant. A second quadratic invariant is

$$H(y_1, y_2, y_3) = \frac{1}{2}\Big(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} \Big),$$

which represents the kinetic energy.

Inspired by the cover page of Marsden & Ratiu (1999), we present in Fig. 1.1 the sphere with some of the solutions of (1.4) corresponding to $I_1 = 2$, $I_2 = 1$ and $I_3 = 2/3$. They lie on the intersection of the sphere with the ellipsoid given by $H(y_1, y_2, y_3) = Const$. In the left picture we have included the numerical solution (30 steps) obtained by the implicit midpoint rule with step size $h = 0.3$ and initial value $y_0 = (\cos(1.1), 0, \sin(1.1))^T$. It stays exactly on a solution curve. This follows from the fact that the implicit midpoint rule preserves quadratic invariants exactly (Sect. IV.2).

For the explicit Euler method (right picture of Fig. 1.1, 320 steps with $h = 0.05$ and the same initial value) we see that the numerical solution shows a wrong qualitative behaviour (it should lie on a closed curve). The numerical solution even drifts away from the sphere.

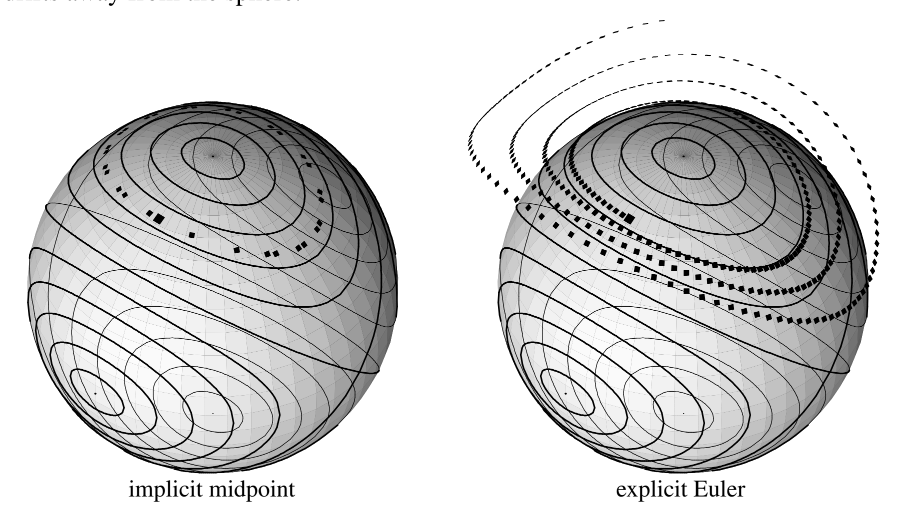


Fig. 1.1. Solutions of the Euler equations (1.4) for the rigid body

IV.2 Quadratic Invariants

Quadratic invariants appear often in applications. Examples are the conservation law of angular momentum in N-body systems (Example 1.3), the two invariants of the rigid body motion (Example 1.7), and the invariant $Y^T Y$ of Theorem 1.6. We therefore consider differential equations (1.1) and quadratic functions

$$Q(y) = y^T C y, \tag{2.1}$$

where C is a symmetric square matrix. It is an invariant of (1.1) if $y^T C f(y) = 0$ for all y.

IV.2.1 Runge–Kutta Methods

We shall give a complete characterization of Runge–Kutta methods which automatically conserve all quadratic invariants. We first of all consider the Gauss collocation methods.

Theorem 2.1. *The Gauss methods of Sect. II.1.3 (collocation based on the shifted Legendre polynomials) conserve quadratic invariants.*

Proof. Let $u(t)$ be the collocation polynomial of the Gauss methods (Definition II.1.3). Since $\frac{d}{dt}Q\big(u(t)\big) = 2u(t)^T C\dot{u}(t)$, it follows from $u(t_0) = y_0$ and $u(t_0 + h) = y_1$ that

$$y_1^T C y_1 - y_0^T C y_0 = 2\int_{t_0}^{t_0+h} u(t)^T C\dot{u}(t)\,dt. \tag{2.2}$$

The integrand $u(t)^T C\dot{u}(t)$ is a polynomial of degree $2s-1$, which is integrated without error by the s-stage Gaussian quadrature formula. It therefore follows from the collocation condition

$$u(t_0 + c_i h)^T C\dot{u}(t_0 + c_i h) = u(t_0 + c_i h)^T C f\big(u(t_0 + c_i h)\big) = 0$$

that the integral in (2.2) vanishes. □

Since the implicit midpoint rule is the special case $s = 1$ of the Gauss methods, the preceding theorem explains its good behaviour for the rigid body simulation in Fig 1.1.

Theorem 2.2 (Cooper 1987). *If the coefficients of a Runge–Kutta method satisfy*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \qquad \text{for all } \ i, j = 1, \ldots, s, \tag{2.3}$$

then it conserves quadratic invariants.[2]

[2] For irreducible methods, the conditions of Theorem 2.2 and Theorem 2.4 are also necessary for the conservation of all quadratic invariants. This follows from the discussion in Sect. VI.7.3.

Proof. The proof is the same as that for B-stability, given independently by Burrage & Butcher and Crouzeix in 1979 (see Hairer & Wanner (1996), Sect. IV.12).

The relation $y_1 = y_0 + h\sum_{i=1}^s b_i k_i$ of Definition II.1.1 yields

$$y_1^T C y_1 = y_0^T C y_0 + h\sum_{i=1}^s b_i\, k_i^T C y_0 + h\sum_{j=1}^s b_j\, y_0^T C k_j + h^2 \sum_{i,j=1}^s b_i b_j\, k_i^T C k_j. \tag{2.4}$$

We then write $k_i = f(Y_i)$ with $Y_i = y_0 + h\sum_{j=1}^s a_{ij}k_j$. The main idea is to compute y_0 from this relation and to insert it into the central expressions of (2.4). This yields (using the symmetry of C)

$$y_1^T C y_1 = y_0^T C y_0 + 2h\sum_{i=1}^s b_i\, Y_i^T C f(Y_i) + h^2 \sum_{i,j=1}^s (b_i b_j - b_i a_{ij} - b_j a_{ji})\, k_i^T C k_j.$$

The condition (2.3) together with the assumption $y^T C f(y) = 0$, which states that $y^T C y$ is an invariant of (1.1), imply $y_1^T C y_1 = y_0^T C y_0$. □

The criterion (2.3) is very restrictive. One finds that among all collocation and discontinuous collocation methods (Definition II.1.7) only the Gauss methods satisfy this criterion (Exercise 6). On the other hand, it is possible to construct other high-order Runge–Kutta methods satisfying (2.3). The key for such a construction is the W-transformation (see Hairer & Wanner (1996), Sect. IV.5), which is exploited in the articles of Sun (1993a) and Hairer & Leone (2000).

IV.2.2 Partitioned Runge–Kutta Methods

We next consider partitioned Runge–Kutta methods for systems $\dot y = f(y,z)$, $\dot z = g(y,z)$. Usually such methods cannot conserve general quadratic invariants (Exercise 4). We therefore concentrate on quadratic invariants of the form

$$Q(y,z) = y^T D z, \tag{2.5}$$

where D is a matrix of the appropriate dimensions. Observe that the angular momentum of N-body systems (Example 1.3) is of this form.

Theorem 2.3 (Sun 1993b). *The Lobatto IIIA - IIIB pair conserves all quadratic invariants of the form (2.5). In particular, this is true for the Störmer–Verlet scheme (see Sect. II.2.2).*

Proof. Let $u(t)$ and $v(t)$ be the (discontinuous) collocation polynomials of the Lobatto IIIA and Lobatto IIIB methods, respectively (see Sect. II.2.2). In analogy to the proof of Theorem 2.1 we have

$$\begin{aligned} &Q\big(u(t_0+h), v(t_0+h)\big) - Q\big(u(t_0), v(t_0)\big) \\ &\qquad = \int_{t_0}^{t_0+h} \Big(Q\big(\dot u(t), v(t)\big) + Q\big(u(t), \dot v(t)\big)\Big)\, dt. \end{aligned} \tag{2.6}$$

Since $u(t)$ is of degree s and $v(t)$ of degree $s-2$, the integrand of (2.6) is a polynomial of degree $2s-3$. Hence, an application of the Lobatto quadrature yields the exact result. Using the fact that $Q(y,z)$ is an invariant of the differential equation, i.e., $Q\big(f(y,z),z\big)+Q\big(y,g(y,z)\big)\equiv 0$, we thus obtain for the integral in (2.6)

$$hb_1\,Q\big(u(t_0),\delta(t_0)\big)+hb_s\,Q\big(u(t_0+h),\delta(t_0+h)\big),$$

where $\delta(t)=\dot v(t)-g\big(u(t),v(t)\big)$ denotes the defect. It now follows from $u(t_0)=y_0$, $u(t_0+h)=y_1$ (definition of Lobatto IIIA) and from $v(t_0)=z_0-hb_1\delta(t_0)$, $v(t_0+h)=z_1+hb_s\delta(t_0+h)$ (definition of Lobatto IIIB) that $Q(y_1,z_1)-Q(y_0,z_0)=0$, which proves the theorem. □

Exchanging the role of the IIIA and IIIB methods also leads to an integrator that preserves quadratic invariants of the form (2.5). The following characterization extends Theorem 2.2 to partitioned Runge–Kutta methods.

Theorem 2.4. *If the coefficients of a partitioned Runge–Kutta method (II.2.2) satisfy*

$$b_i\widehat a_{ij}+\widehat b_j a_{ji}\;=\;b_i\widehat b_j \qquad \textit{for } i,j=1,\ldots,s, \tag{2.7}$$

$$b_i\;=\;\widehat b_i \qquad \textit{for } i=1,\ldots,s, \tag{2.8}$$

then it conserves quadratic invariants of the form (2.5).

If the partitioned differential equation is of the special form $\dot y=f(z)$, $\dot z=g(y)$, then condition (2.7) alone implies that invariants of the form (2.5) are conserved.

Proof. The proof is nearly identical to that of Theorem 2.2. Instead of (2.4) we get

$$y_1^TDz_1=y_0^TDz_0+h\sum_{i=1}^s b_i\,k_i^TDz_0+h\sum_{j=1}^s\widehat b_j\,y_0^TD\ell_j+h^2\sum_{i,j=1}^s b_i\widehat b_j\,k_i^TD\ell_j.$$

Denoting by (Y_i,Z_i) the arguments of $k_i=f(Y_i,Z_i)$ and $\ell_i=g(Y_i,Z_i)$, the same trick as in the proof of Theorem 2.2 gives

$$\begin{aligned} y_1^TDz_1\;=\;y_0^TDz_0\;&+\;h\sum_{i=1}^s b_i\,f(Y_i,Z_i)^TDZ_i+h\sum_{j=1}^s\widehat b_j\,Y_j^TDg(Y_j,Z_j)\\ &+\;h^2\sum_{i,j=1}^s\big(b_i\widehat b_j-b_i\widehat a_{ij}-\widehat b_j a_{ji}\big)\,k_i^TD\ell_j. \end{aligned} \tag{2.9}$$

Since (2.5) is an invariant, we have $f(y,z)^TDz+y^TDg(y,z)=0$ for all y and z. Consequently, the two conditions (2.7) and (2.8) imply $y_1^TDz_1=y_0^TDz_0$.

For the special case where f depends only on z and g only on y, the assumption $f(z)^TDz+y^TDg(y)=0$ (for all y,z) implies that $f(z)^TDz=-y^TDg(y)=Const$. Therefore, condition (2.8) is no longer necessary for the proof of the statement. □

IV.2.3 Nyström Methods

An important class of partitioned differential equations is $\dot{y} = z$, $\dot{z} = g(y)$ or, equivalently,

$$\ddot{y} = g(y). \tag{2.10}$$

Many examples of Chap. I are of this form, in particular the N-body problem of Example 1.3 for which the angular momentum is a quadratic first integral. Nyström methods (Definition II.2.3),

$$\begin{aligned} \ell_i &= g\Big(y_0 + c_i h\, \dot{y}_0 + h^2 \sum_{j=1}^{s} a_{ij}\ell_j\Big), \\ y_1 &= y_0 + h\dot{y}_0 + h^2 \sum_{i=1}^{s} \beta_i \ell_i, \qquad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^{s} b_i \ell_i, \end{aligned} \tag{2.11}$$

are adapted to the numerical solution of (2.10) and it is interesting to investigate which methods within this class can conserve quadratic invariants.

Theorem 2.5. *If the coefficients of the Nyström method (2.11) satisfy*

$$\begin{aligned} \beta_i &= b_i(1 - c_i) \qquad \textit{for } i = 1, \dots, s, \\ b_i(\beta_j - a_{ij}) &= b_j(\beta_i - a_{ji}) \qquad \textit{for } i, j = 1, \dots, s, \end{aligned} \tag{2.12}$$

then it conserves all quadratic invariants of the form $y^T D\, \dot{y}$.

Proof. The quadratic form $Q(y, \dot{y}) = y^T D\, \dot{y}$ is a first integral of (2.10) if and only if

$$\dot{y}^T D\, \dot{y} + y^T D\, g(y) = 0 \qquad \text{for all } y, \dot{y} \in \mathbb{R}^n. \tag{2.13}$$

This implies that D is skew-symmetric and that $y^T D\, g(y) = 0$.

In the same way as for the proofs of Theorems 2.2 and 2.4 we now compute $y_1^T D\, \dot{y}_1$ using the formulas of (2.11) and we substitute y_0 by $Y_i - c_i h \dot{y}_0 - h^2 \sum_j a_{ij}\ell_j$, where Y_i denotes the argument of g in (2.11). This yields

$$\begin{aligned} y_1^T D\, \dot{y}_1 &= y_0^T D\, \dot{y}_0 + h\, \dot{y}_0^T D\, \dot{y}_0 + h \sum_{i=1}^{s} b_i\, Y_i^T D\, \ell_i \\ &+ h^2 \sum_{i=1}^{s} \beta_i\, \ell_i^T D\, \dot{y}_0 + h^2 \sum_{i=1}^{s} b_i(1 - c_i)\, \dot{y}_0^T D\, \ell_i \\ &+ h^3 \sum_{i,j=1}^{s} b_i(\beta_j - a_{ij})\, \ell_j^T D\, \ell_i. \end{aligned}$$

Using the skew-symmetry of D and $Y_i^T D\, \ell_i = Y_i^T D\, g(Y_i) = 0$, condition (2.12) implies the conservation property $y_1^T D\, \dot{y}_1 = y_0^T D\, \dot{y}_0$. □

Remark 2.6 (Composition Methods). If a method Φ_h conserves quadratic invariants (e.g., the mid-point rule by Theorem 2.1 or the Störmer–Verlet scheme by Theorem 2.3 or a Nyström method of Theorem 2.5), then so does the composition method

$$\Psi_h = \Phi_{\gamma_s h} \circ \ldots \circ \Phi_{\gamma_1 h}. \tag{2.14}$$

This obvious property is one of the most important motivations for considering composition methods.

IV.3 Polynomial Invariants

We consider two classes of problems with polynomial invariants for degree higher than two. First, we treat linear problems for which the determinant of the resolvent is an invariant, and we show that (partitioned) Runge–Kutta methods cannot conserve them automatically. Second, we study isospectral flows.

IV.3.1 The Determinant as a First Integral

We consider quasi-linear problems

$$\dot{Y} = A(Y)Y, \qquad Y(0) = Y_0 \tag{3.1}$$

where Y and $A(Y)$ are $n \times n$ matrices. In the following we denote the trace of a matrix $A = (a_{ij})_{i,j=1}^n$ by $\operatorname{trace} A = \sum_{i=1}^n a_{ii}$.

Lemma 3.1. *If* $\operatorname{trace} A(Y) = 0$ *for all* Y*, then* $g(Y) := \det Y$ *is an invariant of the matrix differential equation (3.1).*

Proof. It follows from

$$\det\bigl(Y + \varepsilon AY\bigr) = \det\bigl(I + \varepsilon A\bigr) \det Y = \bigl(1 + \varepsilon \operatorname{trace} A + \mathcal{O}(\varepsilon^2)\bigr) \det Y$$

that $g'(Y)(AY) = \operatorname{trace} A \cdot \det Y$ (this is the *Abel–Liouville–Jacobi–Ostrogradskii identity*). Hence, the determinant $g(Y) = \det Y$ is an invariant of the differential equation (3.1) if $\operatorname{trace} A(Y) = 0$ for all Y. □

Since $\det Y$ represents the volume of the parallelepiped generated by the columns of the matrix Y, the conservation of the invariant $g(Y) = \det Y$ is related to volume preservation. This topic will be further discussed in Sect. VI.9. Here, we consider $\det Y$ as a polynomial invariant of degree n, and we investigate whether Runge–Kutta methods can automatically conserve this invariant for $n \geq 3$. The key lemma for this study is the following.

Lemma 3.2 (Feng Kang & Shang Zai-jiu 1995). *Let $R(z)$ be a differentiable function defined in a neighbourhood of $z = 0$, and assume that $R(0) = 1$ and $R'(0) = 1$. Then, we have for $n \geq 3$*

$$\det R(A) = 1 \quad \textit{for all } n \times n \textit{ matrices } A \textit{ satisfying } \operatorname{trace} A = 0, \tag{3.2}$$

if and only if $R(z) = \exp(z)$.

Proof. The "if" part follows from Lemma 3.1, because for constant A the solution of $\dot{Y} = AY, Y(0) = I$ is given by $Y(t) = \exp(At)$.

For the proof of the "only if" part, we consider diagonal matrices of the form $A = \operatorname{diag}(\mu, \nu, -(\mu + \nu), 0, \ldots, 0)$, which have $\operatorname{trace} A = 0$, and for which

$$R(A) = \operatorname{diag}\Big(R(\mu), R(\nu), R(-(\mu + \nu)), R(0), \ldots, R(0)\Big).$$

The assumptions $R(0) = 1$ and (3.2) imply

$$R(\mu)R(\nu)R(-(\mu + \nu)) = 1 \tag{3.3}$$

for all μ, ν close to 0. Putting $\nu = 0$, this relation yields $R(\mu)R(-\mu) = 1$ for all μ, and therefore (3.3) can be written as

$$R(\mu)R(\nu) = R(\mu + \nu) \qquad \text{for all } \mu, \nu \text{ close to } 0. \tag{3.4}$$

This functional equation can only be satisfied by the exponential function. This is seen as follows: from (3.4) we have

$$\frac{R(\mu + \varepsilon) - R(\mu)}{\varepsilon} = R(\mu)\, \frac{R(\varepsilon) - R(0)}{\varepsilon}.$$

Taking the limit $\varepsilon \to 0$ we obtain $R'(\mu) = R(\mu)$, because $R'(0) = 1$. This implies $R(\mu) = \exp(\mu)$. □

Theorem 3.3. *For $n \geq 3$, no Runge–Kutta method can conserve all polynomial invariants of degree n.*

Proof. It is sufficient to consider linear problems $\dot{Y} = AY$ with constant matrix A satisfying $\operatorname{trace} A = 0$, so that $g(Y) = \det Y$ is a polynomial invariant of degree n. Applying a Runge–Kutta method to such a differential equation yields $Y_1 = R(hA)Y_0$, where

$$R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}$$

($b^T = (b_1, \ldots, b_s)$, $\mathbb{1} = (1, \ldots, 1)^T$ and $\mathcal{A} = (a_{ij})$ is the matrix of Runge–Kutta coefficients) is the so-called stability function. It is seen to be rational. By Lemma 3.2 it is therefore not possible that $\det R(hA) = 1$ for all A with $\operatorname{trace} A = 0$. □

This negative result motivates the search for new methods which can conserve polynomial invariants (see Sects. IV.4, IV.8 and VI.9). We consider here another interesting class of problems with polynomial invariants of degree higher than two.

IV.3.2 Isospectral Flows

Such flows are created by a matrix differential equation

$$\dot{L} = [B(L), L], \qquad L(0) = L_0 \tag{3.5}$$

where L_0 is a given symmetric matrix, $B(L)$ is skew-symmetric for all L, and $[B, L] = BL - LB$ is the commutator of B and L. Many interesting problems can be written in this form. We just mention the Toda system, the continuous realization of QR-type algorithms, projected gradient flows, and inverse eigenvalue problems (see Chu (1992) and Calvo, Iserles & Zanna (1997) for long lists of references).

Lemma 3.4 (Lax 1968, Flaschka 1974). *Let L_0 be symmetric and assume that $B(L)$ is skew-symmetric for all L. Then, the solution $L(t)$ of (3.5) is a symmetric matrix, and its eigenvalues are independent of t.*

Proof. The symmetry of $L(t)$ follows from the fact that the commutator of a skew-symmetric with a symmetric matrix gives a symmetric matrix.

To prove the isospectrality of the flow, we define $U(t)$ by

$$\dot{U} = B\big(L(t)\big)\, U, \qquad U(0) = I. \tag{3.6}$$

Then, we have $(d/dt)(U^{-1}LU) = U^{-1}(\dot{L} - BL + LB)U = 0$, and hence $U(t)^{-1}L(t)U(t) = L_0$ for all t, so that $L(t) = U(t)L_0U(t)^{-1}$ is the solution of (3.5). This proves the result. □

Note that, since $B(L)$ is skew-symmetric, the matrix $U(t)$ of (3.6) is orthogonal by Theorem 1.6.

Lemma 3.4 shows that the characteristic polynomial $\det(L-\lambda I) = \sum_{i=0}^{n} a_i\lambda^i$ and hence the coefficients a_i also are independent of t. These coefficients are all polynomial invariants (e.g., $a_0 = \det L$, $a_{n-1} = \pm\operatorname{trace} L$). Because of Theorem 3.3 there is no hope that Runge–Kutta methods applied to (3.5) can conserve these invariants automatically for $n \geq 3$.

Isospectral Methods. The proof of Lemma 3.4, however, suggests an interesting approach for the numerical solution of (3.5). For $n = 0, 1, \ldots$ we solve numerically

$$\dot{U} = B(UL_nU^T)U, \qquad U(0) = I \tag{3.7}$$

and we put $L_{n+1} = \widehat{U}L_n\widehat{U}^T$, where $\widehat{U}$ is the numerical approximation $\widehat{U} \approx U(h)$ after one step (cf. Calvo, Iserles & Zanna 1999). If $B(L)$ is skew-symmetric for all matrices L, then U^TU is a quadratic invariant of (3.7) and the methods of Sect. IV.2 will produce an orthogonal $\widehat{U}$. Consequently, L_{n+1} and L_n have exactly the same eigenvalues, and they remain symmetric.

Diele, Lopez & Politi (1998) suggest the use of the Cayley transform $U = (I-Y)^{-1}(I+Y)$, which transforms (3.7) into

$$\dot{Y} = \frac{1}{2}(I-Y)B\big(UL_nU^T\big)(I+Y), \qquad Y(0) = 0,$$

and the orthogonality of U into the skew-symmetry of Y (see Lemma 8.8 below). Since all (also explicit) Runge–Kutta methods preserve the skew-symmetry of Y, which is a linear invariant, this yields an approach to explicit isospectral methods.

Connection with the QR Algorithm. In a diversion from the main theme of this section, we now show the relationship of the flow of (3.5) with the QR algorithm for the symmetric eigenvalue problem. Starting from a real symmetric matrix A_0, the basic *QR algorithm* (without shifts) computes a sequence of orthogonally similar matrices $A_1, A_2, A_3, \ldots$, expected to converge towards a diagonal matrix carrying the eigenvalues of A_0. Iteratively for $k = 0, 1, 2, \ldots$, one computes the QR decomposition of A_k:

$$A_k = Q_k R_k$$

with Q_k orthogonal, R_k upper triangular (the decomposition becomes unique if the diagonal elements of R_k are taken positive). Then, A_{k+1} is obtained by reversing the order of multiplication:

$$A_{k+1} = R_k Q_k.$$

It is an easy exercise to show that $Q(k) = Q_0 Q_1 \ldots Q_{k-1}$ is the matrix in the orthogonal similarity transformation between A_0 and A_k:

$$A_k = Q(k)^T A_0 Q(k) \tag{3.8}$$

and the same matrix $Q(k)$ is the orthogonal factor in the QR decomposition of A_0^k:

$$A_0^k = Q(k)R(k). \tag{3.9}$$

Consider now, for an arbitrary real function f defined on the eigenvalues of a real symmetric matrix L_0, the QR decomposition

$$\exp(tf(L_0)) = Q(t)R(t) \tag{3.10}$$

and define

$$L(t) := Q(t)^T L_0 Q(t). \tag{3.11}$$

The relations (3.8) and (3.9) then show that for integer times $t = k$, the matrix $\exp\big(f(L(k))\big) = Q(k)^T \exp\big(f(L_0)\big)Q(k)$ coincides with the kth matrix in the QR algorithm starting from $A_0 = \exp(f(L_0))$:

$$\exp(f(L(k))) = A_k\ . \tag{3.12}$$

Now, how is all this related to the system (3.5)? Differentiating (3.11) as in the proof of Lemma 3.4 shows that $L(t)$ solves a differential equation of the form $\dot{L} = [B, L]$ with the skew-symmetric matrix $B = -Q^T \dot{Q}$. At first sight, however, B is a function of t, not of L. On the other hand, differentiation of (3.10) yields (omitting the argument t where it is clear from the context)

$$f(L_0)QR = f(L_0)\exp(tf(L_0)) = \exp(tf(L_0))f(L_0) = \dot{Q}R + Q\dot{R}\ ,$$

and since $f(L) = Q^T f(L_0) Q$ by (3.11), this becomes

$$f(L) = Q^T \dot{Q} + \dot{R} R^{-1} .$$

Here the left-hand side is a symmetric matrix, and the right-hand side is the sum of a skew-symmetric and an upper triangular matrix. It follows that the skew-symmetric matrix $B = -Q^T \dot{Q}$ is given by

$$B(L) = f(L)_+ - f(L)_+^T, \tag{3.13}$$

where $f(L)_+$ denotes the part of $f(L)$ above the diagonal. Hence, $L(t)$ is the solution of an autonomous system (3.5) with a skew-symmetric $B(L)$.

For $f(x) = x$ and assuming L_0 symmetric and tridiagonal, the flow of (3.5) with (3.13) is known as the *Toda flow*. The QR iterates $A_0 = \exp(L_0)$, $A_1, A_2, \ldots$ of the exponential of L_0 are seen to be equal to the exponentials of the solution $L(t)$ of the Toda equations at integer times: $A_k = \exp(L(k))$, a discovery of Symes (1982). An interesting connection of the Toda equations with a mechanical system will be discussed in Sect. X.1.5.

For $f(x) = \log x$, the above arguments show that the QR iteration itself, starting from a positive definite symmetric tridiagonal matrix, is the evaluation $A_k = L(k)$ at integer times of a solution $L(t)$ of the differential equation (3.5) with B given by (3.13). This relationship was explored in a series of papers by Deift, Li, Nanda & Tomei (1983, 1989, 1993).

Notwithstanding the mathematical beauty of this relationship, it must be remarked that the practical QR algorithm (with shifts and deflation) follows a different path.

IV.4 Projection Methods

Und bist du nicht willig, so brauch ich Gewalt.
(J.W. Goethe, *Der Erlkönig*)

Suppose we have an $(n-m)$-dimensional submanifold of $\mathbb{R}^n$,

$$\mathcal{M} = \{y \; ; \; g(y) = 0\} \tag{4.1}$$

($g : \mathbb{R}^n \to \mathbb{R}^m$), and a differential equation $\dot{y} = f(y)$ with the property that

$$y_0 \in \mathcal{M} \qquad \text{implies} \qquad y(t) \in \mathcal{M} \;\text{ for all } t. \tag{4.2}$$

We want to emphasize that this assumption is weaker than the requirement that all components $g_i(y)$ of $g(y)$ are invariants in the sense of Definition 1.1. In fact, assumption (4.2) is equivalent to $g'(y)f(y) = 0$ for $y \in \mathcal{M}$, whereas Definition 1.1 requires $g'(y)f(y) = 0$ for all $y \in \mathbb{R}^n$. In the situation of (4.2) we call $g(y)$ a *weak invariant*, and we say that $\dot{y} = f(y)$ is a differential equation on the manifold $\mathcal{M}$.

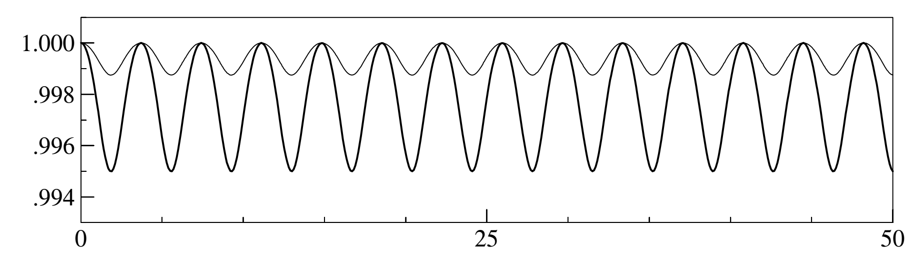


Fig. 4.1. The implicit midpoint rule applied to the differential equation (4.3). The picture shows the numerical values for $q_1^2 + q_2^2$ obtained with step size $h = 0.1$ (thick line) and $h = 0.05$ (thin line)

Example 4.1. Consider the pendulum equation written in Cartesian coordinates:

$$\begin{aligned} \dot q_1 &= p_1, \qquad & \dot p_1 &= -q_1\lambda, \\ \dot q_2 &= p_2, \qquad & \dot p_2 &= -1 - q_2\lambda, \end{aligned} \tag{4.3}$$

where $\lambda = (p_1^2 + p_2^2 - q_2)/(q_1^2 + q_2^2)$. One can check by differentiation that $q_1p_1 + q_2p_2$ (orthogonality of the position and velocity vectors) is an invariant in the sense of Definition 1.1. However, $q_1^2+q_2^2$ (length of the pendulum) is only a weak invariant. The experiment of Fig. 4.1 shows that even methods which conserve quadratic first integrals (cf. Sect. IV.2) do not conserve the quadratic weak invariant $q_1^2 + q_2^2$. No numerical method that is allowed to evaluate the vector field $f(y)$ outside $\mathcal{M}$ can be expected to conserve weak invariants exactly. This is one of the motivations for considering the methods of this and the subsequent sections.

A natural approach to the numerical solution of differential equations on manifolds is by projection (see e.g., Hairer & Wanner (1996), Sect. VII.2, Eich-Soellner & Führer (1998), Sect. 5.3.3).

Algorithm 4.2 (Standard Projection Method). *Assume that $y_n \in \mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.2):*

- *Compute $\widetilde y_{n+1} = \Phi_h(y_n)$, where Φ_h is an arbitrary one-step method applied to $\dot y = f(y)$;*
- *project the value $\widetilde y_{n+1}$ onto the manifold $\mathcal{M}$ to obtain $y_{n+1} \in \mathcal{M}$.*

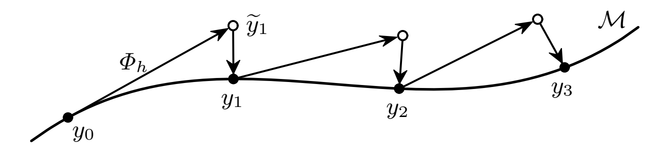


Fig. 4.2. Illustration of the standard projection method

For $y_n \in \mathcal{M}$ the distance of $\widetilde y_{n+1}$ to the manifold $\mathcal{M}$ is of the size of the local error, i.e., $\mathcal{O}(h^{p+1})$. Therefore, the projection does not deteriorate the convergence order of the method.

For the computation of y_{n+1} we have to solve the constrained minimization problem

$$\|y_{n+1} - \widetilde{y}_{n+1}\| \to \min \qquad \text{subject to} \qquad g(y_{n+1}) = 0. \tag{4.4}$$

In the case of the Euclidean norm, a standard approach is to introduce Lagrange multipliers $\lambda = (\lambda_1, \dots, \lambda_m)^T$, and to consider the Lagrange function $\mathcal{L}(y_{n+1}, \lambda) = \|y_{n+1} - \widetilde{y}_{n+1}\|^2/2 - g(y_{n+1})^T\lambda$. The necessary condition $\partial\mathcal{L}/\partial y_{n+1} = 0$ then leads to the system

$$\begin{aligned} y_{n+1} &= \widetilde{y}_{n+1} + g'(\widetilde{y}_{n+1})^T\lambda \\ 0 &= g(y_{n+1}). \end{aligned} \tag{4.5}$$

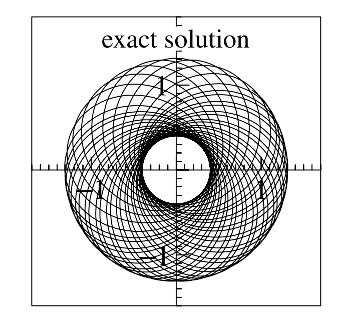
We have replaced y_{n+1} with $\widetilde{y}_{n+1}$ in the argument of $g'(y)$ in order to save some evaluations of $g'(y)$. Inserting the first relation of (4.5) into the second gives a nonlinear equation for λ, which can be efficiently solved by simplified Newton iterations:

$$\Delta\lambda_i = -\Big(g'(\widetilde{y}_{n+1})g'(\widetilde{y}_{n+1})^T\Big)^{-1} g\Big(\widetilde{y}_{n+1} + g'(\widetilde{y}_{n+1})^T\lambda_i\Big), \qquad \lambda_{i+1} = \lambda_i + \Delta\lambda_i.$$

For the choice $\lambda_0 = 0$ the first increment $\Delta\lambda_0$ is of size $\mathcal{O}(h^{p+1})$, so that the convergence is usually extremely fast. Often, one simplified Newton iteration is sufficient.

Example 4.3. As a first example we consider the perturbed Kepler problem (see Exercise I.12) with Hamiltonian function

$$\begin{aligned} H(p,q) &= \frac{1}{2}\left(p_1^2 + p_2^2\right) - \frac{1}{\sqrt{q_1^2 + q_2^2}} \\ &- \frac{0.005}{2\sqrt{(q_1^2 + q_2^2)^3}}, \end{aligned}$$



and initial values $q_1(0) = 1 - e$, $q_2(0) = 0$, $p_1(0) = 0$, $p_2(0) = \sqrt{(1+e)/(1-e)}$ (eccentricity $e = 0.6$) on the interval $0 \le t \le 200$. The exact solution (plotted to the right) is approximately an ellipse that rotates slowly around one of its foci. For this problem we know two first integrals: the Hamiltonian function $H(p,q)$ and the angular momentum $L(p,q) = q_1p_2 - q_2p_1$.

We apply the explicit Euler method and the symplectic Euler method (I.1.9), both with constant step size $h = 0.03$. The result is shown in Fig. 4.3. The numerical solution of the explicit Euler method (without projection) is completely wrong. The projection onto the manifold $\{H(p,q) = H(p_0,q_0)\}$ improves the numerical solution, but it still has a wrong qualitative behaviour. Only projection onto both invariants, $H(p,q) = Const$ and $L(p,q) = Const$ gives the correct behaviour. The symplectic Euler method already shows the correct behaviour without any projections (see Chap. IX for an explanation). Surprisingly, a projection onto $H(p,q) = Const$ destroys this behaviour, the numerical solution approaches the centre and the simplified Newton iterations fail to converge beyond $t = 25.23$. Projection onto both invariants re-establishes the correct behaviour.

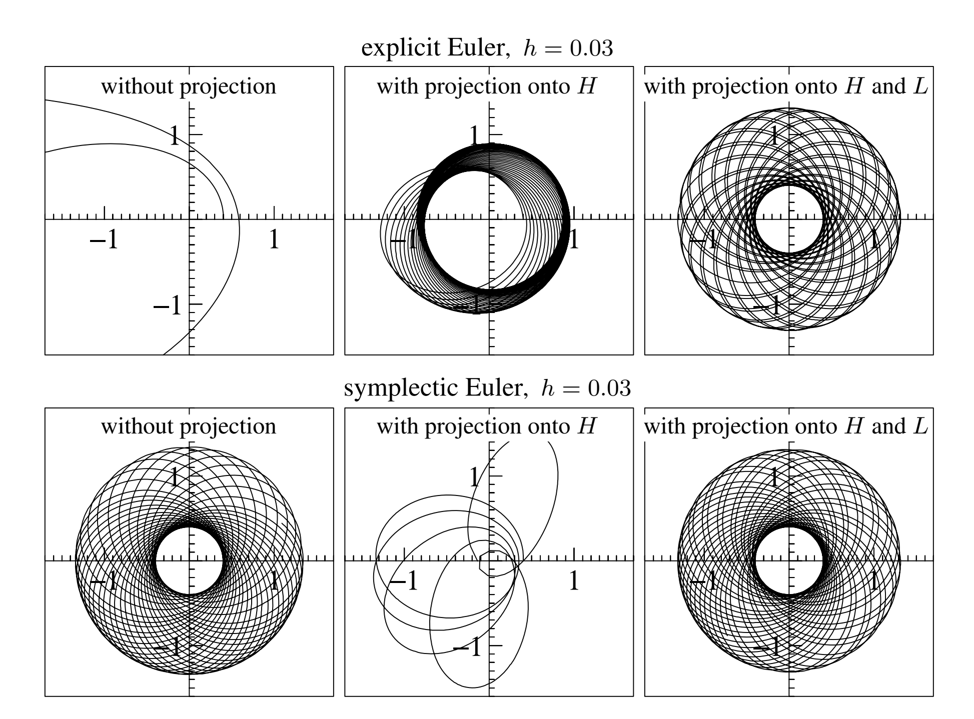


Fig. 4.3. Numerical solutions obtained with and without projections

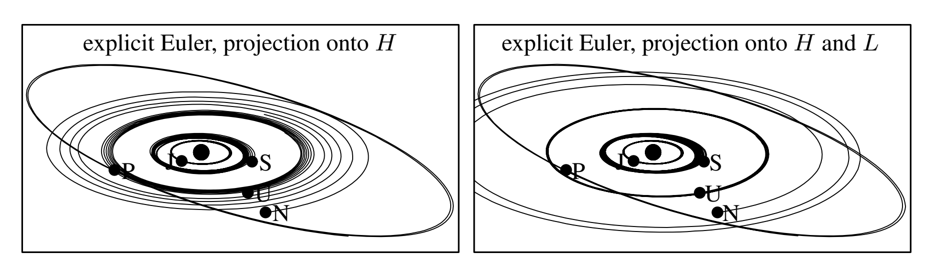


Fig. 4.4. Explicit Euler method with projections applied to the outer solar system, step size $h = 10$ (days), interval $0 \le t \le 200\,000$

Example 4.4 (Outer Solar System). Having encountered excellent experience with projections onto H and L for the perturbed Kepler problem (Example 4.3), let us apply the same idea to a more realistic problem in celestial mechanics. We consider the outer solar system as described in Sect. I.2. The numerical solution of the explicit Euler method applied with constant step size $h = 10$, once with projection onto $H = Const$ and once with projection onto $H = Const$ and $L = Const$, is shown in Fig. 4.4 (observe that the conservation of the angular momentum $L(p,q) = \sum_{i=1}^{N} q_i \times p_i$ consists of three first integrals). We see a slight improvement in the orbits of Jupiter, Saturn and Uranus (compared to the explicit

Euler method without projections, see Fig. I.2.4), but the orbit of Neptune becomes even worse. There is no doubt that this problem contains a structure which cannot be correctly simulated by methods that only preserve the total energy H and the angular momentum L.

Example 4.5 (Volume Preservation). Consider the matrix differential equation $\dot{Y} = A(Y)Y$, where $\operatorname{trace} A(Y) = 0$ for all Y. We know from Lemma 3.1 that $g(Y) = \det Y$ is an invariant which cannot be automatically conserved by Runge–Kutta methods. Here, we show how we can enforce this invariant by projection. Let $\widetilde{Y}_{n+1}$ be the numerical approximation obtained with an arbitrary one-step method. We consider the Frobenius norm $\|Y\|_F = \sqrt{\sum_{i,j} |y_{ij}|^2}$ for measuring the distance to the manifold $\{Y\,;\, g(Y) = 0\}$. Using $g'(Y)(AY) = \operatorname{trace} A \, \det Y$ (see the proof of Lemma 3.1) with A chosen such that the product AY contains only one non-zero element, the projection step (4.5) is seen to become (Exercise 9)

$$Y_{n+1} = \widetilde{Y}_{n+1} + \mu \widetilde{Y}_{n+1}^{-T} \tag{4.6}$$

with the scalar $\mu = \lambda \det \widetilde{Y}_{n+1}$. This leads to the scalar nonlinear equation $\det\big(\widetilde{Y}_{n+1} + \mu \widetilde{Y}_{n+1}^{-T}\big) = \det Y_n$, for which simplified Newton iterations become

$$\det\Big(\widetilde{Y}_{n+1} + \mu_i \widetilde{Y}_{n+1}^{-T}\Big)\Big(1 + (\mu_{i+1} - \mu_i)\operatorname{trace}\big((\widetilde{Y}_{n+1}^T \widetilde{Y}_{n+1})^{-1}\big)\Big) = \det Y_n.$$

If the QR-decomposition of $\widetilde{Y}_{n+1}$ is available from the computation of $\det \widetilde{Y}_{n+1}$, the value of $\operatorname{trace}\big((\widetilde{Y}_{n+1}^T \widetilde{Y}_{n+1})^{-1}\big)$ can be computed efficiently with $\mathcal{O}(n^3/3)$ flops (see e.g., Golub & Van Loan (1989), Sect. 5.3.9).

The above projection is preferable to $Y_{n+1} = c\widetilde{Y}_{n+1}$, where $c \in \mathbb{R}$ is chosen such that $\det Y_{n+1} = \det Y_n$. This latter projection is already ill-conditioned for diagonal matrices with entries that differ by several magnitudes.

As a conclusion to the above numerical experiments we see that a projection can give excellent results, but can also destroy the good long-time behaviour of the solution if applied inappropriately. If the original method already preserves some structure, then projection to a subset of invariants may destroy the good long-time behaviour. An important modification for reversible differential equations (symmetric projections) will be presented in Sect. V.4.1.

IV.5 Numerical Methods Based on Local Coordinates

A second important class of methods for the numerical treatment of differential equations on manifolds uses local coordinates. Before explaining the ideas, we find it appropriate to discuss in more detail manifolds and differential equations on manifolds.

IV.5.1 Manifolds and the Tangent Space

In Sect. IV.4 we assumed that locally (in a neighbourhood U of $a \in \mathbb{R}^n$) a manifold is given by constraints, i.e.,

$$\mathcal{M} = \{y \in U \; ; \; g(y) = 0\}, \tag{5.1}$$

where $g : U \to \mathbb{R}^m$ is differentiable, $g(a) = 0$, and $g'(a)$ has full rank m.

Here, we use local parameters to characterize a manifold. Let $\psi : V \to \mathbb{R}^n$ be differentiable ($V \subset \mathbb{R}^{n-m}$ is a neighbourhood of 0), $\psi(0) = a$, and assume that $\psi'(0)$ has full rank $n - m$. Then, a manifold is locally given by

$$\mathcal{M} = \{y = \psi(z) \; ; \; z \in V\} \tag{5.2}$$

provided that V is sufficiently small, so that $\psi : V \to \psi(V)$ is bijective with continuous inverse. The variables z are called *parameters* or *local coordinates* of the manifold.

As an example, consider the unit sphere which, in the form (5.1), is given by the function $g(y_1, y_2, y_3) = y_1^2 + y_2^2 + y_3^2 - 1$. There are many possible choices of local coordinates. Away from the equator (i.e., $y_3 = 0$), we can take $z = (z_1, z_2)^T := (y_1, y_2)^T$ and $\psi(z) = \big(z_1, z_2, \pm\sqrt{1 - z_1^2 - z_2^2}\big)^T$. Alternatively, we can consider spherical coordinates $\psi(\alpha, \beta) = \big(\cos\alpha \sin\beta, \sin\alpha \sin\beta, \cos\beta\big)^T$ away from the north and south poles (i.e., $y_1 = y_2 = 0, y_3 = \pm 1$).

The tangent to a curve (or the tangent plane to a surface) is an affine space passing through the contact point $a \in \mathcal{M}$. It is convenient to place the origin at a, so that we obtain a vector space. More precisely, for a manifold $\mathcal{M}$ we define the *tangent space* at $a \in \mathcal{M}$ as

$$T_a\mathcal{M} = \left\{ v \in \mathbb{R}^n \,\middle|\, \begin{array}{l} \text{there exists a differentiable path } \gamma : (-\varepsilon, \varepsilon) \to \mathbb{R}^n \\ \text{with } \; \gamma(t) \in \mathcal{M} \text{ for all } t, \; \gamma(0) = a, \; \dot\gamma(0) = v \end{array} \right\}. \tag{5.3}$$

Lemma 5.1. *If the manifold $\mathcal{M}$ is given by (5.1), where $g : U \to \mathbb{R}^m$ is differentiable, $g(a) = 0$, and $g'(a)$ has full rank m, then we have*

$$T_a\mathcal{M} = \ker g'(a) = \{v \in \mathbb{R}^n \,|\, g'(a)v = 0\}. \tag{5.4}$$

If $\mathcal{M}$ is given by (5.2), where $\psi : V \to \mathbb{R}^n$ is differentiable, $\psi(0) = a$, and $\psi'(0)$ has full rank $n - m$, then we have

$$T_a\mathcal{M} = \mathrm{Im}\, \psi'(0) = \{\psi'(0)w \,|\, w \in \mathbb{R}^{n-m}\}. \tag{5.5}$$

Proof. a) For a path $\gamma(t)$ satisfying $\gamma(0) = a$ and $g\big(\gamma(t)\big) = 0$ it follows by differentiation that $g'(a)\dot\gamma(0) = 0$. Consequently, we have $T_a\mathcal{M} \subset \ker g'(a)$.

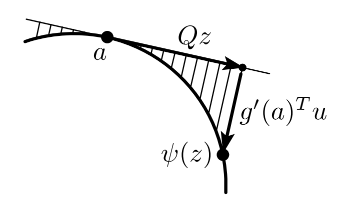
Consider now the function $F(t, u) = g\big(a + tv + g'(a)^T u\big)$. We have $F(0, 0) = 0$ and an invertible $\partial F/\partial u(0, 0) = g'(a)g'(a)^T$, so that by the implicit function theorem the relation $F(t, u) = 0$ can be solved locally for $u = u(t)$. If $v \in \ker g'(a)$,

it follows that $\dot{u}(0) = 0$, and the path $\gamma(t) = a + tv + g'(a)^T u(t)$ satisfies all requirements of (5.3), so that also $T_a\mathcal{M} \supset \ker g'(a)$.

b) Assume $\mathcal{M}$ to be given by (5.2). For an arbitrary $\eta : (-\varepsilon, \varepsilon) \to \mathbb{R}^m$ satisfying $\eta(0) = 0$, the path $\gamma(t) = \psi\big(\eta(t)\big)$ lies in $\mathcal{M}$ and satisfies $\dot\gamma(0) = \psi'(0)\dot\eta(0)$. This proves $\operatorname{Im}\psi'(0) \subset T_a\mathcal{M}$.

The assumption on the rank of $\psi'(0)$ implies that, after a reordering of the components, we have $\psi(z) = (\psi_1(z), \psi_2(z))^T$, where $\psi_1(z)$ is a local diffeomorphism (by the inverse function theorem). We show that every smooth path $\gamma(t)$ in $\mathcal{M}$ can be written as $\gamma(t) = \psi\big(\eta(t)\big)$ with some smooth $\eta(t)$. This then implies $T_a\mathcal{M} \subset \operatorname{Im}\psi'(0)$. To prove this we split $\gamma(t) = (\gamma_1(t), \gamma_2(t))^T$ according to the partitioning of ψ, and we define $\eta(t) = \psi_1^{-1}\big(\gamma_1(t)\big)$. Since for $\gamma(t) \in \mathcal{M}$ the second part $\gamma_2(t)$ is uniquely determined by $\gamma_1(t)$, this proves $\gamma(t) = \psi\big(\eta(t)\big)$. □

The proof of the preceding lemma shows the equivalence of the representations (5.1) and (5.2) of manifolds in $\mathbb{R}^n$. Let $\mathcal{M}$ be given by (5.1), and assume that the columns of Q form an orthogonal basis of $T_a\mathcal{M}$. As in part (a) of the proof of Lemma 5.1 the condition $g\big(a + Qz + g'(a)^T u\big) = 0$ defines locally (close to $z = 0$) a function $u(z)$ which satisfies $u(0) = 0$ and $u'(0) = 0$. Hence, the manifold $\mathcal{M}$ is also given by (5.2) with the function $\psi(z) = a + Qz + g'(a)^T u(z)$.



On the other hand, let $\mathcal{M}$ be given by (5.2). Part (b) of the proof of Lemma 5.1 shows that $y = \psi(z)$ can be partitioned into $y_1 = \psi_1(z)$ and $y_2 = \psi_2(z)$, where ψ_1 is a local diffeomorphism. Consequently, $\mathcal{M}$ is also given by (5.1) with $g(y) = y_2 - \psi_2\big(\psi_1^{-1}(y_1)\big)$.

IV.5.2 Differential Equations on Manifolds

In Sect. IV.4 we introduced differential equations on a manifold as problems satisfying (4.2). With the help of Lemma 5.1 we are now in a position to characterize such problems without knowledge of the solutions.

Theorem 5.2. *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^n$. The problem $\dot y = f(y)$ is a differential equation on the manifold $\mathcal{M}$ (i.e., it satisfies (4.2)) if and only if*

$$f(y) \in T_y\mathcal{M} \qquad \textit{for all } y \in \mathcal{M}. \tag{5.6}$$

Proof. The necessity of (5.6) follows from the definition of $T_y\mathcal{M}$, because the exact solution of the differential equation lies in $\mathcal{M}$ and has $f(y)$ as derivative.

To prove the sufficiency, we assume (5.6) and let $\mathcal{M}$ be locally, near y_0, be given by a parametrization $y = \psi(z)$ as in (5.2). We try to write the solution of $\dot y = f(y)$, $y(0) = y_0 = \psi(z_0)$ as $y(t) = \psi(z(t))$. If this is at all possible, then $z(t)$ must satisfy

$$\psi'(z)\dot{z} = f\big(\psi(z)\big)$$

which, by assumption (5.6) and the second part of Lemma 5.1, is equivalent to

$$\dot{z} = \psi'(z)^+ f\big(\psi(z)\big), \tag{5.7}$$

where $A^+ = (A^T A)^{-1} A^T$ denotes the pseudo-inverse of a matrix with full column rank. Conversely, define $z(t)$ as the solution of (5.7) with $z(0) = z_0$, which is known to exist locally in t by the standard existence and uniqueness theory of ordinary differential equations on $\mathbb{R}^m$. Then $y(t) = \psi(z(t))$ is the solution of $\dot{y} = f(y)$ with $y(0) = y_0$. Hence, the solution $y(t)$ remains in $\mathcal{M}$. □

We remark that the sufficiency proof of Theorem 5.2 only requires the function $f(y)$ to be defined on $\mathcal{M}$. Due to the equivalence of $\dot{y} = f(y)$ with (5.7) the problem is transported to the space of local coordinates. The standard local theory for ordinary differential equations on an Euclidean space (existence and uniqueness of solutions, ...) can thus be extended in a straightforward way to differential equations on manifolds, i.e., $\dot{y} = f(y)$ with $f : \mathcal{M} \to \mathbb{R}^n$ satisfying (5.6).

IV.5.3 Numerical Integrators on Manifolds

Whereas the projection methods of Sect. IV.4 require the function $f(y)$ of the differential equation to be defined in a neighbourhood of $\mathcal{M}$ (see Fig. 4.2), the numerical methods of this section evaluate $f(y)$ only on the manifold $\mathcal{M}$. The idea is to apply the numerical integrator in the parameter space rather than in the space where $\mathcal{M}$ is embedded.

Algorithm 5.3 (Local Coordinates Approach). *Assume that $y_n \in \mathcal{M}$ and that ψ is a local parametrization of $\mathcal{M}$ satisfying $\psi(z_n) = y_n$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 5.1):*

- *Compute $\widetilde{z}_{n+1} = \Phi_h(z_n)$, the result of the method Φ_h applied to (5.7);*
- *define the numerical solution by $y_{n+1} = \psi(\widetilde{z}_{n+1})$.*

It is important to remark that the parametrization $y = \psi(z)$ can be changed at every step.

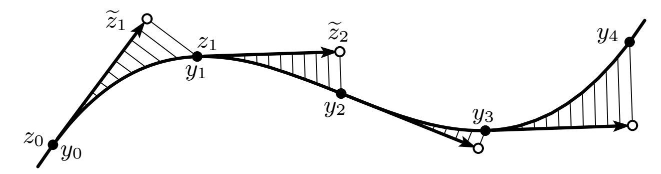


Fig. 5.1. The numerical solution of differential equations on manifolds via local coordinates

As indicated at the beginning of Sect. IV.5.1, there are many possible choices of local coordinates. Consider the pendulum equation of Example 4.1, where $\mathcal{M} = \{(q_1, q_2, p_1, p_2) \mid q_1^2 + q_2^2 = 1,\ q_1p_1 + q_2p_2 = 0\}$. A standard parametrization here is $q_1 = \sin\alpha$, $q_2 = -\cos\alpha$, $p_1 = \omega\cos\alpha$, and $p_2 = \omega\sin\alpha$. In the new coordinates (α, ω) the problem becomes simply $\dot\alpha = \omega$, $\dot\omega = -\sin\alpha$. Other typical choices are the exponential map $\psi(Z) = \exp(Z)$ for differential equations on Lie groups, and the Cayley transform $\psi(Z) = (I - Z)^{-1}(I + Z)$ for quadratic Lie groups. This will be studied in more detail in Sect. IV.8 below. Here we discuss two commonly used choices which do not use a special structure of the manifold.

Generalized Coordinate Partitioning. We assume that the manifold is given by (5.1). If $g : \mathbb{R}^n \to \mathbb{R}^m$ has a Jacobian with full rank m at $y = a$, we can find a partitioning $y = (y_1, y_2)$, such that $\partial g/\partial y_2(a)$ is invertible. In this case we can choose the components of y_1 as local coordinates. The function $y = \psi(z)$ is then given by $y_1 = z$ and $y_2 = \psi_2(z)$, where $\psi_2(z)$ is implicitly defined by $g\big(z, \psi_2(z)\big) = 0$. This approach has been promoted by Wehage & Haug (1982) in the context of constrained mechanical systems, and the partitioning is found by Gaussian elimination with full pivoting applied to the matrix $g'(a)$. Another way of finding the partitioning is by the use of the QR decomposition with column change.

Tangent Space Parametrization. Let the manifold $\mathcal{M}$ be given by (5.1), and collect the vectors of an orthogonal basis of $T_a\mathcal{M}$ in the matrix Q. We then consider the parametrization

$$\psi_a(z) = a + Qz + g'(a)^T u(z), \tag{5.8}$$

where $u(z)$ is defined by $g\big(\psi_a(z)\big) = 0$, exactly as in the discussion after the proof of Lemma 5.1. Differentiating (5.8) yields

$$\big(Q + g'(a)^T u'(z)\big)\dot z = \dot y = f(y) = f\big(\psi_a(z)\big).$$

Since $Q^TQ = I$ and $g'(a)Q = 0$, this relation is equivalent to the differential equation

$$\dot z = Q^T f\big(\psi_a(z)\big), \tag{5.9}$$

which corresponds to (5.7). If we apply a numerical method to (5.9), every function evaluation requires the projection of an element of the tangent space onto the manifold. This procedure is illustrated in Fig. 5.1, and was originally proposed by Potra & Rheinboldt (1991) for the solution of the Euler–Lagrange equations of constrained multibody systems (see also Hairer & Wanner (1996), p. 476).

IV.6 Differential Equations on Lie Groups

Theorem 1.6 and Lemma 3.1 are particular cases of a more general result which can be conveniently formulated with the concept of Lie groups and Lie algebras (see Olver (1986) and Varadarajan (1974) for an introduction to these subjects).

A *Lie group* is a group G which is a differentiable manifold, and for which the product is a differentiable mapping $G \times G \to G$. We restrict our considerations to *matrix Lie groups*, that is, Lie groups which are subgroups of $\mathrm{GL}(n)$, the group of invertible $n \times n$ matrices with the usual matrix product as the group operation.



Marius Sophus Lie[3]

Example 6.1. An important example of a Lie group is the group

$$\mathrm{O}(n) = \big\{ Y \in \mathrm{GL}(n) \mid Y^T Y = I \big\}$$

of all orthogonal matrices. It is the zero set of $g(Y) = Y^T Y - I$, where we consider g as a mapping from the set of all $n \times n$ matrices (i.e., $\mathbb{R}^{n \cdot n}$) to the set of all symmetric matrices (which can be identified with $\mathbb{R}^{n(n+1)/2}$). The derivative $g'(Y)$ is surjective for $Y \in \mathrm{O}(n)$, because for any symmetric matrix K the choice $H = YK/2$ solves the equation $g'(Y)H = K$. Therefore, the matrix $g'(Y)$ has full rank (cf. (5.1)) so that $\mathrm{O}(n)$ defines a differentiable manifold of dimension $n^2 - n(n+1)/2 = n(n-1)/2$. The set $\mathrm{O}(n)$ is also a group with unit element I (the identity). Since the matrix multiplication is a differentiable mapping, $\mathrm{O}(n)$ is a Lie group.

Table 6.1 lists further prominent examples. The matrix J appearing in the definition of the symplectic group is the matrix determining the symplectic structure on $\mathbb{R}^n$ (see Sect. VI.2).

As the following lemma shows, the tangent space $\mathfrak{g} = T_I G$ at the identity I of a matrix Lie group G is closed under forming commutators of its elements. This makes $\mathfrak{g}$ an algebra, the *Lie algebra* of the Lie group G.

Lemma 6.2 (Lie Bracket and Lie Algebra). *Let G be a matrix Lie group and let $\mathfrak{g} = T_I G$ be the tangent space at the identity. The Lie bracket (or commutator)*

$$[A, B] = AB - BA \tag{6.1}$$

defines an operation $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ which is bilinear, skew-symmetric ($[A, B] = -[B, A]$), and satisfies the Jacobi identity

$$\big[A, [B, C]\big] + \big[C, [A, B]\big] + \big[B, [C, A]\big] = 0. \tag{6.2}$$

[3] Marius Sophus Lie, born: 17 December 1842 in Nordfjordeid (Norway), died: 18 February 1899.

Table 6.1. Some matrix Lie groups and their corresponding Lie algebras

Lie group			Lie algebra		
$\mathrm{GL}(n)$	$=$	$\{Y \mid \det Y \neq 0\}$ general linear group	$\mathfrak{gl}(n)$	$=$	$\{A \mid \text{arbitrary matrix}\}$ Lie algebra of $n \times n$ matrices
$\mathrm{SL}(n)$	$=$	$\{Y \mid \det Y = 1\}$ special linear group	$\mathfrak{sl}(n)$	$=$	$\{A \mid \text{trace}(A) = 0\}$ special linear Lie algebra
$\mathrm{O}(n)$	$=$	$\{Y \mid Y^T Y = I\}$ orthogonal group	$\mathfrak{so}(n)$	$=$	$\{A \mid A^T + A = 0\}$ skew-symmetric matrices
$\mathrm{SO}(n)$	$=$	$\{Y \in \mathrm{O}(n) \mid \det Y = 1\}$ special orthogonal group	$\mathfrak{so}(n)$	$=$	$\{A \mid A^T + A = 0\}$ skew-symmetric matrices
$\mathrm{Sp}(n)$	$=$	$\{Y \mid Y^T J Y = J\}$ symplectic group	$\mathfrak{sp}(n)$	$=$	$\{A \mid JA + A^T J = 0\}$

Proof. By definition of the tangent space, for $A, B \in \mathfrak{g}$, there exist differentiable paths $\alpha(t), \beta(t)$ ($|t| < \varepsilon$) in G such that $\alpha(t) = I + tA(t)$ with a continuous function $A(t)$ with $A(0) = A$, and similarly $\beta(t) = I + tB(t)$ with $B(0) = B$. Now consider the path $\gamma(t)$ in G defined by

$$\gamma(t) = \alpha(\sqrt{t})\beta(\sqrt{t})\alpha(\sqrt{t})^{-1}\beta(\sqrt{t})^{-1}, \quad t \geq 0.$$

An elementary computation then yields

$$\gamma(t) = I + t[A, B] + o(t).$$

With the extension $\gamma(t) = \gamma(-t)^{-1}$ for negative t, this is a differentiable path in G satisfying $\gamma(0) = I$ and $\dot\gamma(0) = [A, B]$. Hence $[A, B] \in \mathfrak{g}$ by definition of the tangent space. The properties of the Lie bracket can be verified in a straightforward way. □

Example 6.3. Consider again the orthogonal group $\mathrm{O}(n)$. Since the derivative of $g(Y) = Y^T Y - I$ at the identity is $g'(I)H = I^T H + H^T I = H + H^T$, it follows from the first part of Lemma 5.1 that the Lie algebra corresponding to $\mathrm{O}(n)$ consists of all skew-symmetric matrices. The right column of Table 6.1 gives the Lie algebras of the other Lie groups listed there.

The following basic lemma shows that the exponential map yields a local parametrization of the Lie group near the identity, with the Lie algebra (a linear space) as the parameter space.

Lemma 6.4 (Exponential Map). *Consider a matrix Lie group G and its Lie algebra $\mathfrak{g}$. The matrix exponential is a map*

$$\exp : \mathfrak{g} \to G,$$

i.e., for $A \in \mathfrak{g}$ we have $\exp(A) \in G$. Moreover, $\exp$ is a local diffeomorphism in a neighbourhood of $A = 0$.

Proof. For $A \in \mathfrak{g}$, it follows from the definition of the tangent space $\mathfrak{g} = T_I G$ that there exists a differentiable path $\alpha(t)$ in G satisfying $\alpha(0) = I$ and $\dot\alpha(0) = A$. For a fixed $Y \in G$, the path $\gamma(t) := \alpha(t)Y$ is in G and satisfies $\gamma(0) = Y$ and $\dot\gamma(0) = AY$. Consequently, $AY \in T_Y G$ and $\dot Y = AY$ defines a differential equation on the manifold G. The solution $Y(t) = \exp(tA)$ is therefore in G for all t.

Since $\exp(H) - \exp(0) = H + \mathcal{O}(H^2)$, the derivative of the exponential map at $A = 0$ is the identity, and it follows from the inverse function theorem that $\exp$ is a local diffeomorphism close to $A = 0$. □

The proof of Lemma 6.4 shows that for a matrix Lie group G the tangent space at $Y \in G$ has the form

$$T_Y G = \{AY \mid A \in \mathfrak{g}\}. \tag{6.3}$$

By Theorem 5.2, differential equations on a matrix Lie group (considered as a manifold) can therefore be written as

$$\dot Y = A(Y)Y \tag{6.4}$$

where $A(Y) \in \mathfrak{g}$ for all $Y \in G$. The following theorem summarizes this discussion, and extends the statements of Theorem 1.6 and Lemma 3.1 to more general matrix Lie groups.

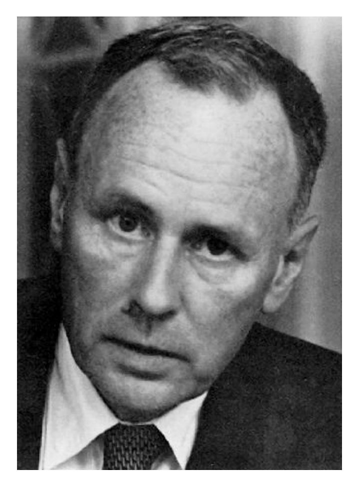
Theorem 6.5. *Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. If $A(Y) \in \mathfrak{g}$ for all $Y \in G$ and if $Y_0 \in G$, then the solution of (6.4) satisfies $Y(t) \in G$ for all t.* □

If in addition $A(Y) \in \mathfrak{g}$ for all matrices Y, and if

$$G = \{Y \mid g(Y) = Const\}$$

is one of the Lie groups of Table 6.1, then $g(Y)$ is an invariant of the differential equation (6.4) in the sense of Definition 1.1.

IV.7 Methods Based on the Magnus Series Expansion



Wilhelm Magnus[4]

Before we discuss the numerical solution of differential equations (6.4) on Lie groups, let us give an explicit formula for the solution of linear matrix differential equations

$$\dot{Y} = A(t)Y. \tag{7.1}$$

No assumption on the matrix $A(t)$ is made for the moment (apart from continuous dependence on t). For the scalar case, the solution of (7.1) with $Y(0) = Y_0$ is given by

$$Y(t) = \exp\Big(\int_0^t A(\tau)\,d\tau\Big)Y_0. \tag{7.2}$$

Also in the case where the matrices $A(t)$ and $\int_0^t A(\tau)\,d\tau$ commute, (7.2) is the solution of (7.1). In the general non-commutative case we follow the approach of Magnus (1954) and we search for a matrix function $\Omega(t)$ such that

$$Y(t) = \exp\big(\Omega(t)\big)Y_0$$

solves (7.1). The main ingredient for the solution will be the inverse of the derivative of the matrix exponential. It has been studied in Sect. III.4, Lemma III.4.2, and is given by

$$d\exp_\Omega^{-1}(H) = \sum_{k\ge 0} \frac{B_k}{k!}\,\mathrm{ad}_\Omega^k(H), \tag{7.3}$$

where B_k are the Bernoulli numbers, and $\mathrm{ad}_\Omega(A) = [\Omega, A] = \Omega A - A\Omega$ is the adjoint operator introduced in (III.4.1).

Theorem 7.1 (Magnus 1954). *The solution of the differential equation (7.1) can be written as $Y(t) = \exp\big(\Omega(t)\big)Y_0$ with $\Omega(t)$ defined by*

$$\dot{\Omega} = d\exp_\Omega^{-1}\big(A(t)\big), \qquad \Omega(0) = 0. \tag{7.4}$$

As long as $\|\Omega(t)\| < \pi$, the convergence of the $d\exp_\Omega^{-1}$ expansion (7.3) is assured.

Proof. Comparing the derivative of $Y(t) = \exp\big(\Omega(t)\big)Y_0$,

$$\dot{Y}(t) = \Big(\frac{d}{d\Omega}\exp\Omega(t)\Big)\,\dot{\Omega}(t)Y_0 = \Big(d\exp_{\Omega(t)}\big(\dot{\Omega}(t)\big)\Big)\,\exp\big(\Omega(t)\big)Y_0,$$

with (7.1) we obtain $A(t) = d\exp_{\Omega(t)}\big(\dot{\Omega}(t)\big)$. Applying the inverse operator $d\exp_\Omega^{-1}$ to this relation yields the differential equation (7.4) for $\Omega(t)$. The statement on the convergence is a consequence of Lemma III.4.2. □

[4] Wilhelm Magnus, born: 5 February 1907 in Berlin (Germany), died: 15 October 1990.

The first few Bernoulli numbers are $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$. The differential equation (7.4) therefore becomes

$$\dot{\Omega} = A(t) - \frac{1}{2}\,[\Omega, A(t)] + \frac{1}{12}\Big[\Omega, [\Omega, A(t)]\Big] + \ldots\,,$$

which is nonlinear in Ω. Applying Picard fixed point iteration after integration yields

$$\begin{aligned}\Omega(t) \;=\; &\int_0^t A(\tau)\,d\tau \;-\; \frac{1}{2}\int_0^t \Big[\int_0^\tau A(\sigma)\,d\sigma, A(\tau)\Big]\,d\tau \\ &+\; \frac{1}{4}\int_0^t \Big[\int_0^\tau \Big[\int_0^\sigma A(\mu)\,d\mu, A(\sigma)\Big]\,d\sigma, A(\tau)\Big]\,d\tau \\ &+\; \frac{1}{12}\int_0^t \Big[\int_0^\tau A(\sigma)\,d\sigma, \Big[\int_0^\tau A(\mu)\,d\mu, A(\tau)\Big]\Big]\,d\tau + \ldots\,,\end{aligned} \tag{7.5}$$

which is the so-called *Magnus expansion*. For smooth matrices $A(t)$ the remainder in (7.5) is of size $\mathcal{O}(t^5)$ so that the truncated series inserted into $Y(t) = \exp\big(\Omega(t)\big)Y_0$ gives an excellent approximation to the solution of (7.1) for small t.

Numerical Methods Based on the Magnus Expansion. Iserles & Nørsett (1999) study the general form of the Magnus expansion (7.5), and they relate the iterated integrals and the rational coefficients in (7.5) to binary trees. For a numerical integration of

$$\dot{Y} = A(t)Y, \qquad Y(t_0) = Y_0 \tag{7.6}$$

(where Y is a matrix or a vector) they propose using $Y_{n+1} = \exp(h\Omega_n)Y_n$, where $h\Omega_n$ is a suitable approximation of $\Omega(h)$ given by (7.5) with $A(t_n + \tau)$ instead of $A(\tau)$. Of course, the Magnus expansion has to be truncated and the integrals have to be approximated by numerical quadrature.

We follow here the collocation approach suggested by Zanna (1999). The idea is to replace $A(t)$ locally by an interpolation polynomial

$$\widehat{A}(t) = \sum_{i=1}^{s} \ell_i(t)\,A(t_n + c_i h),$$

and to solve $\dot{Y} = \widehat{A}(t)Y$ on $[t_n, t_n + h]$ by the use of the truncated series (7.5).

Theorem 7.2. *Consider a quadrature formula $(b_i, c_i)_{i=1}^s$ of order $p \ge s$, and let $Y(t)$ and $Z(t)$ be solutions of $\dot{Y} = A(t)Y$ and $\dot{Z} = \widehat{A}(t)Z$, respectively, satisfying $Y(t_n) = Z(t_n)$. Then, $Z(t_n + h) - Y(t_n + h) = \mathcal{O}(h^{p+1})$.*

Proof. We write the differential equation for Z as $\dot{Z} = A(t)Z + \big(\widehat{A}(t) - A(t)\big)Z$ and use the variation of constants formula to get

$$Z(t_n + h) - Y(t_n + h) = \int_{t_n}^{t_n+h} R(t_n + h, \tau)\Big(\widehat{A}(\tau) - A(\tau)\Big)Z(\tau)\,d\tau.$$

Applying our quadrature formula to this integral gives zero as result, and the remainder is of size $\mathcal{O}(h^{p+1})$. Details of the proof are as for Theorem II.1.5. □

Example 7.3. As a first example, we use the midpoint rule ($c_1 = 1/2$, $b_1 = 1$). In this case the interpolation polynomial is constant, and the method becomes

$$Y_{n+1} = \exp\Big(hA(t_n + h/2)\Big)\, Y_n, \tag{7.7}$$

which is of order 2.

Example 7.4. The two-stage Gauss quadrature is given by $c_{1,2} = 1/2 \pm \sqrt{3}/6$, $b_{1,2} = 1/2$. The interpolation polynomial is of degree one and we have to apply (7.5) in order to get an approximation Y_{n+1}. Since we are interested in a fourth order approximation, we can neglect the remainder term (indicated by ... in (7.5)). Computing analytically the iterated integrals over products of $\ell_i(t)$ we obtain

$$Y_{n+1} = \exp\Big(\frac{h}{2}(A_1 + A_2) + \frac{\sqrt{3}\,h^2}{12}[A_2, A_1]\Big)\, Y_n, \tag{7.8}$$

where $A_1 = A(t_n + c_1 h)$ and $A_2 = A(t_n + c_2 h)$. This is a method of order four. The terms of (7.5) with triple integrals give $\mathcal{O}(h^4)$ expressions, whose leading term vanishes by the symmetry of the method (Exercise V.7). Therefore, they need not be considered.

Theorem 7.2 allows us to obtain methods of arbitrarily high order. A straightforward use of the expansion (7.5) yields an expression with a large number of commutators. Munthe-Kaas & Owren (1999) and Blanes, Casas & Ros (2000a) construct higher order methods with a reduced number of commutators. For example, for order 6 the required number of commutators is reduced from 7 to 4.

Let us remark that all numerical methods of this section are of the form $Y_{n+1} = \exp(h\Omega_n)Y_n$, where Ω_n is a linear combination of $A(t_n + c_i h)$ and of their commutators. If $A(t) \in \mathfrak{g}$ for all t, then also $h\Omega_n$ lies in the Lie algebra $\mathfrak{g}$, so that the numerical solution stays in the Lie group G if $Y_0 \in G$ (this is a consequence of Lemma 6.4).

IV.8 Lie Group Methods

Consider a differential equation

$$\dot{Y} = A(Y)Y, \qquad Y(0) = Y_0 \tag{8.1}$$

on a matrix Lie group G. This means that $Y_0 \in G$ and that $A(Y) \in \mathfrak{g}$ for all $Y \in G$. Since this is a special case of differential equations on a manifold, projection methods (Sect. IV.4) as well as methods based on local coordinates (Sect. IV.5) are well suited for their numerical treatment. Here we present further approaches which also yield approximations that lie on the manifold.

All numerical methods of this section can be extended in a straightforward way to non-autonomous problems $\dot{Y} = A(t, Y)Y$ with $A(t, Y) \in \mathfrak{g}$ for all t and all $Y \in G$. Just to simplify the notation we restrict ourselves to the formulation (8.1).

IV.8.1 Crouch-Grossman Methods

> The discipline of Lie-group methods owes a great deal to the pioneering work of Peter Crouch and his co-workers . . .
> (A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett & A. Zanna 2000)

The numerical approximation of explicit Runge–Kutta methods is obtained by a composition of the following two basic operations: (i) an evaluation of the vector field $f(Y) = A(Y)Y$ and (ii) a computation of an update of the form $Y + haf(Z)$. For example, the left method of (II.1.3) consists of the following steps: evaluate $K_1 = f(Y_0)$; compute $\widetilde{Y}_1 = Y_0 + hK_1$; evaluate $K_2 = f(\widetilde{Y}_1)$; compute $Y_{1/2} = Y_0 + \frac{h}{2}K_1$; compute $Y_1 = Y_{1/2} + \frac{h}{2}K_2$.

In the context of differential equations on Lie groups, these methods have the disadvantage that, even when $Y \in G$ and $Z \in G$, the update $Y + haA(Z)Z$ is in general not in the Lie group. The idea of Crouch & Grossman (1993) is to replace the "update" operation with $\exp\big(haA(Z)\big)Y$.

Definition 8.1. Let b_i, a_{ij} $(i,j = 1,\dots,s)$ be real numbers. An explicit *s-stage Crouch-Grossman method* is given by

$$\begin{aligned} Y^{(i)} &= \exp(ha_{i,i-1}K_{i-1}) \cdot \ldots \cdot \exp(ha_{i1}K_1)Y_n, \qquad K_i = A(Y^{(i)}), \\ Y_{n+1} &= \exp(hb_sK_s) \cdot \ldots \cdot \exp(hb_1K_1)Y_n. \end{aligned}$$

For example, the method of Runge described above ($s = 2$, $a_{21} = 1$, $b_1 = b_2 = 1/2$) leads to

$$Y_{n+1} = \exp\Big(\frac{h}{2}K_2\Big)\exp\Big(\frac{h}{2}K_1\Big)Y_n, \tag{8.2}$$

where $K_1 = A(Y_n)$ and $K_2 = A\big(\exp(hK_1)Y_n\big)$.

By construction, the methods of Crouch-Grossman give rise to approximations Y_n which lie exactly on the manifold defined by the Lie group. But what can be said about their order of accuracy?

Theorem 8.2. *Let $c_i = \sum_j a_{ij}$. A Crouch-Grossman method has order p ($p \le 3$) if the following order conditions are satisfied:*

$$\begin{aligned} &\textit{order } 1: && \textstyle\sum_i b_i = 1 && (8.3)\\ &\textit{order } 2: && \textstyle\sum_i b_ic_i = 1/2 && (8.4)\\ &\textit{order } 3: && \textstyle\sum_i b_ic_i^2 = 1/3 && (8.5)\\ & && \textstyle\sum_{ij} b_ia_{ij}c_j = 1/6 && (8.6)\\ & && \textstyle\sum_i b_i^2c_i + 2\sum_{i<j} b_ic_ib_j = 1/3. && (8.7) \end{aligned}$$

Proof. As in the case of Runge–Kutta methods, the order conditions can be found by comparing the Taylor series expansions of the exact and the numerical solution. In addition to the conditions stated in the theorem, this leads to relations such as

$$\sum_i b_i^2 c_i + 2 \sum_{i<j} b_i b_j c_j = \frac{2}{3}. \tag{8.8}$$

Adding this equation to (8.7) we find $2\sum_{ij} b_i c_i b_j = 1$, which is satisfied by (8.3) and (8.4). Hence, the relation (8.8) is already a consequence of the conditions stated in the theorem. □

Table 8.1. Crouch-Grossman methods of order 3

c	A		
0			
$-1/24$	$-1/24$		
$17/24$	$161/24$	-6	
	1	$-2/3$	$2/3$

c	A		
0			
$3/4$	$3/4$		
$17/24$	$119/216$	$17/108$	
	$13/51$	$-2/3$	$24/17$

Crouch & Grossman (1993) present several solutions of the system (8.3)–(8.7), one of which is given in the left array of Table 8.1. The construction of higher order Crouch-Grossman methods is very complicated ("... any attempt to analyze algorithms of order greater than three will be very complex, ...", Crouch & Grossman, 1993).

The theory of order conditions for Runge–Kutta methods (Sect. III.1) has been extended to Crouch-Grossman methods by Owren & Marthinsen (1999). It turns out that the order conditions for classical Runge–Kutta methods form a subset of those for Crouch-Grossman methods. The first new condition is (8.7). For a method of order 4, thirteen conditions (including those of Theorem 8.2) have to be satisfied. Solving these equations, Owren & Marthinsen (1999) construct a 4th order method with $s = 5$ stages.

IV.8.2 Munthe-Kaas Methods

These methods were developed in a series of papers by Munthe-Kaas (1995, 1998, 1999). The main motivation behind the first of these papers was to develop a theory of Runge–Kutta methods in a coordinate-free framework. After attempts that led to new order conditions (as for the Crouch-Grossman methods), Munthe-Kaas (1999) had the idea to write the solution as $Y(t) = \exp\big(\Omega(t)\big)Y_0$ and to solve numerically the differential equation for $\Omega(t)$. It sounds awkward to replace the differential equation (8.1) by a more complicated one. However, the nonlinear invariants $g(Y) = 0$ of (8.1) defining the Lie group are replaced with linear invariants $g'(I)(\Omega) = 0$ defining the Lie algebra, and we know from Sect. IV.1 that essentially all numerical methods automatically conserve linear invariants.

It follows from the proof of Theorem 7.1 that the solution of (8.1) can be written as $Y(t) = \exp\big(\Omega(t)\big)Y_0$, where $\Omega(t)$ is the solution of $\dot\Omega = d\exp_\Omega^{-1}\big(A\big(Y(t)\big)\big)$, $\Omega(0) = 0$. Since it is not practical to work with the operator $d\exp_\Omega^{-1}$, we truncate the series (7.3) suitably and consider the differential equation

$$\dot{\Omega} = A\big(\exp(\Omega)Y_0\big) + \sum_{k=1}^{q} \frac{B_k}{k!}\,\mathrm{ad}_{\Omega}^{k}\Big(A\big(\exp(\Omega)Y_0\big)\Big), \qquad \Omega(0) = 0. \tag{8.9}$$

This leads to the following method.

Algorithm 8.3 (Munthe-Kaas 1999). *Consider the problem (8.1) with $A(Y) \in \mathfrak{g}$ for $Y \in G$. Assume that Y_n lies in the Lie group G. Then, the step $Y_n \mapsto Y_{n+1}$ is defined as follows:*

- *consider the differential equation (8.9) with Y_n instead of Y_0, and apply a Runge–Kutta method (explicit or implicit) to get an approximation $\Omega_1 \approx \Omega(h)$,*
- *then define the numerical solution by $Y_{n+1} = \exp\big(\Omega_1\big)Y_n$.*

Before analyzing this algorithm, we emphasize its close relationship with Algorithm 5.3. In fact, if we identify the Lie algebra $\mathfrak{g}$ with $\mathbb{R}^k$ (where k is the dimension of the vector space $\mathfrak{g}$), the mapping $\psi(\Omega) = \exp(\Omega)Y_n$ is a local parametrization of the Lie group G (see Lemma 6.4). Apart from the truncation of the series in (8.9), Algorithm 8.3 is a special case of Algorithm 5.3.

Important properties of the Munthe-Kaas methods are given in the next two theorems.

Theorem 8.4. *Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. If $A(Y) \in \mathfrak{g}$ for $Y \in G$ and if $Y_0 \in G$, then the numerical solution of the Lie group method of Algorithm 8.3 lies in G, i.e., $Y_n \in G$ for all $n = 0, 1, 2, \ldots$.*

Proof. It is sufficient to prove that for $Y_0 \in G$ the numerical solution Ω_1 of the Runge–Kutta method applied to (8.9) lies in $\mathfrak{g}$. Since the Lie bracket $[\Omega, A]$ is an operation $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, and since $\exp(\Omega)Y_0 \in G$ for $\Omega \in \mathfrak{g}$, the right-hand expression of (8.9) is in $\mathfrak{g}$ for $\Omega \in \mathfrak{g}$. Hence, (8.9) is a differential equation on the vector space $\mathfrak{g}$ with solution $\Omega(t) \in \mathfrak{g}$. All operations in a Runge–Kutta method give results in $\mathfrak{g}$, so that the numerical approximation Ω_1 also lies in $\mathfrak{g}$. □

Theorem 8.5. *If the Runge–Kutta method is of (classical) order p and if the truncation index in (8.9) satisfies $q \geq p - 2$, then the method of Algorithm 8.3 is of order p.*

Proof. For sufficiently smooth $A(Y)$ we have $\Omega(t) = tA(Y_0) + \mathcal{O}(t^2)$, $Y(t) = Y_0 + \mathcal{O}(t)$ and $[\Omega(t), A(Y(t))] = \mathcal{O}(t^2)$. This implies that $\mathrm{ad}_{\Omega(t)}^{k}(A(Y(t))) = \mathcal{O}(t^{k+1})$, so that the truncation of the series in (8.9) induces an error of size $\mathcal{O}(h^{q+2})$ for $|t| \leq h$. Hence, for $q + 2 \geq p$, this truncation does not affect the order of convergence. □

The most simple Lie group method is obtained if we take the explicit Euler method as basic discretization and $q = 0$ in (8.9). This leads to the so-called *Lie–Euler method*

$$Y_{n+1} = \exp\big(hA(Y_n)\big)Y_n. \tag{8.10}$$

This is also a special case of the Crouch-Grossman methods of Definition 8.1.

Taking the implicit midpoint rule as the basic discretization and again $q = 0$ in (8.9), we obtain the *Lie midpoint rule*

$$Y_{n+1} = \exp(\Omega)Y_n, \qquad \Omega = hA\big(\exp(\Omega/2)Y_n\big). \tag{8.11}$$

This is an implicit equation in Ω and has to be solved by fixed point iteration or by Newton-type methods.

Example 8.6. We take the coefficients of the right array of Table 8.1. They give rise to 3rd order Munthe-Kaas and 3rd order Crouch-Grossman methods. We apply both methods with the large step size $h = 0.35$ to the system (1.5) which is already of the form (8.1). Observe that Y_0 is a vector in $\mathbb{R}^3$ and not a matrix, but all results of this section remain valid for this case. For the computation of the matrix exponential we use the Rodrigues formula (Exercise 17). The numerical results (first 1000 steps) are shown in Fig. 8.1. We see that the numerical solution stays on the manifold (sphere), but on the sphere the qualitative behaviour is not correct. A similar behaviour could be observed for projection methods (the orthogonal projection consists simply in dividing the approximation $\widetilde{Y}_{n+1}$ by its norm) and by the methods based on local coordinates.

Crouch-Grossman methods and Munthe-Kaas methods are very similar. If they are based on the same set of Runge–Kutta coefficients, both methods use s evaluations of the matrix $A(Y)$. The Crouch-Grossman methods require in general the computation of $s(s+1)/2$ matrix exponentials, whereas the Munthe-Kaas methods require only s of them. On the other hand, Munthe-Kaas methods need also the computations of a certain number of commutators which increases with q in (8.9). In such a comparison one has to take into account that every classical Runge–Kutta method defines a Munthe-Kaas method of the same order, but Crouch-Grossman methods of high order are very difficult to obtain, and need more stages for the same order (if $p \geq 4$).

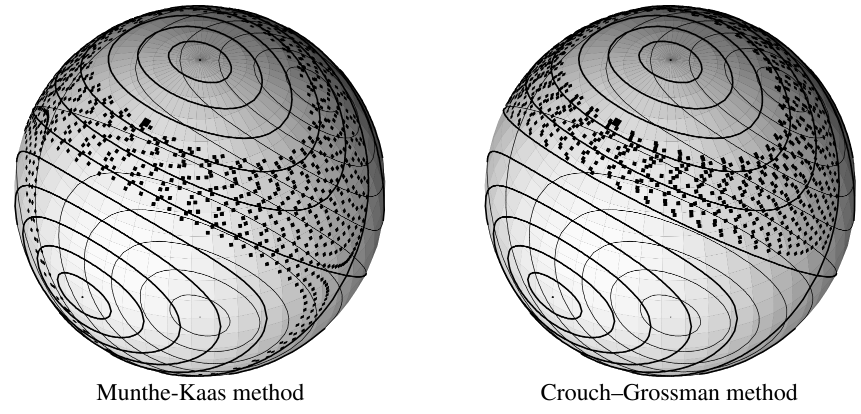


Fig. 8.1. Solutions of the Euler equations (1.4) for the rigid body

IV.8.3 Further Coordinate Mappings

The methods of Algorithm 8.3 are based on the local parametrization $\psi(\Omega) = \exp(\Omega)Y_n$. For all Lie groups, this is a diffeomorphism between the Lie group and the corresponding Lie algebra. Are there other, computationally more efficient parametrizations that can be used in special situations?

The Cayley Transform. Lie groups of the form

$$G = \{Y \mid Y^T P Y = P\}, \tag{8.12}$$

where P is a given constant matrix, are called *quadratic Lie groups*. The corresponding Lie algebra is given by $\mathfrak{g} = \{\Omega \mid P\Omega + \Omega^T P = 0\}$. The orthogonal group $\mathrm{O}(n)$ and the symplectic group $\mathrm{Sp}(n)$ are prominent special cases (see Table 6.1). For such groups we have the following analogue of Lemma 6.4.

Lemma 8.7. *For a quadratic Lie group G, the Cayley transform*

$$\mathrm{cay}\, \Omega = (I - \Omega)^{-1}(I + \Omega)$$

maps elements of $\mathfrak{g}$ into G. Moreover, it is a local diffeomorphism near $\Omega = 0$.

Proof. For $\Omega \in \mathfrak{g}$ (i.e., $P\Omega + \Omega^T P = 0$) we have $P(I + \Omega) = (I - \Omega)^T P$ and also $P(I - \Omega)^{-1} = (I + \Omega)^{-T} P$. For $Y = (I - \Omega)^{-1}(I + \Omega)$ this immediately implies $Y^T P Y = P$. □

The use of the Cayley transform for the numerical integration of differential equations on Lie groups has been proposed by Lewis & Simo (1994) and Diele, Lopez & Peluso (1998) for the orthogonal group, and by Lopez & Politi (2001) for general quadratic groups. It is based on the following result, which is an adaptation of Lemma III.4.1 and Lemma III.4.2 to the Cayley transform.

Lemma 8.8. *The derivative of* $\mathrm{cay}\, \Omega$ *is given by*

$$\Big(\frac{d}{d\Omega}\, \mathrm{cay}\, \Omega\Big) H = \Big(d\,\mathrm{cay}_{\,\Omega}(H)\Big)\, \mathrm{cay}\, \Omega,$$

where

$$d\,\mathrm{cay}_{\,\Omega}(H) = 2(I - \Omega)^{-1} H (I + \Omega)^{-1}. \tag{8.13}$$

For the inverse of $d\,\mathrm{cay}_{\,\Omega}$ *we have*

$$d\,\mathrm{cay}_{\,\Omega}^{-1}(H) = \frac{1}{2}(I - \Omega) H (I + \Omega). \tag{8.14}$$

Proof. By the usual rules of calculus we obtain

$$\Big(\frac{d}{d\Omega}\, \mathrm{cay}\, \Omega\Big) H = (I - \Omega)^{-1} H (I - \Omega)^{-1}(I + \Omega) + (I - \Omega)^{-1} H,$$

and a simple algebraic manipulation proves the statements. □

The numerical approach for solving (8.1) in the case of quadratic Lie groups is an adaptation of the Algorithm 8.3. We consider the local parametrization $Y = \psi(\Omega) = \text{cay}\,(\Omega)Y_n$, and we apply one step of a numerical method to the differential equation $\dot\Omega = d\,\text{cay}\,_\Omega^{-1} A\big(\text{cay}\,(\Omega)Y_n\big)$ which, by (8.14), is equivalent to

$$\dot\Omega = \frac{1}{2}(I-\Omega)A\Big(\text{cay}\,(\Omega)Y_n\Big)(I+\Omega).$$

This equation replaces (8.9) in the Algorithm 8.3. Since no truncation of an infinite series is necessary here, this approach is a special case of Algorithm 5.3.

Canonical Coordinates of the Second Kind. For a basis $\{C_1, C_2, \ldots, C_d\}$ of the Lie algebra $\mathfrak{g}$ the coordinates $z_1, \ldots, z_d$ of the local parametrization $\psi(z) = \exp\big(\sum_{i=1}^d z_i C_i\big)$ of the Lie group G are called *canonical coordinates of the first kind*. Here we are interested in the parametrization

$$\psi(z) = \exp\big(z_1C_1\big)\exp\big(z_2C_2\big)\cdot\ldots\cdot\exp\big(z_dC_d\big), \tag{8.15}$$

and we call $z = (z_1, \ldots, z_d)^T$ *canonical coordinates of the second kind* (Varadarajan 1974). The use of these coordinates in connection with the numerical solution of differential equations on Lie groups has been promoted by Celledoni & Iserles (2001) and Owren & Marthinsen (2001). The idea behind this choice is that, due to a sparse structure of the C_i, the computation of $\exp(z_1C_1), \ldots, \exp(z_dC_d)$ may be much cheaper than the computation of $\exp(\sum_i z_iC_i)$.

With the change of coordinates $y = \psi(z)$, the differential equation (8.1) becomes $\psi'(z)\dot z = A\big((\psi(z)\big)\psi(z)$, which is equivalent to

$$\begin{aligned} A\big(\psi(z)\big) &= \sum_{i=1}^d \dot z_i \,\exp(z_1C_1)\cdot\ldots\cdot\exp(z_{i-1}C_{i-1}) \\ &\qquad\qquad \cdot C_i\cdot\exp(-z_{i-1}C_{i-1})\cdot\ldots\cdot\exp(-z_1C_1) \\ &= \sum_{i=1}^d \dot z_i\,\big(F_1\circ\ldots\circ F_{i-1}\big)C_i, \end{aligned}\tag{8.16}$$

where we use the notation $F_jC = \exp(z_jC_j)\,C\exp(-z_jC_j)$ for the linear operator $F_j : \mathfrak{g}\to\mathfrak{g}$; see Exercise 12. We need to compute $\dot z_1, \ldots, \dot z_d$ from (8.16), and this will usually be a computationally expensive task. However, for several Lie algebras and for well chosen bases this can be done very efficiently. The crucial idea is the following: we let $\widehat F_j$ be defined by

$$\widehat F_j C_i = \begin{cases} F_jC_i & \text{if } i > j \\ C_i & \text{if } i \le j, \end{cases} \tag{8.17}$$

and we assume that

$$\big(F_1\circ\ldots\circ F_{i-1}\big)C_i = \big(\widehat F_1\circ\ldots\circ\widehat F_{i-1}\big)C_i, \qquad i = 2, \ldots, d. \tag{8.18}$$

Under this assumption, we have $\big(F_1 \circ \ldots \circ F_{i-1}\big)C_i = \big(\widehat{F}_1 \circ \ldots \circ \widehat{F}_{i-1}\big)C_i = \big(\widehat{F}_1 \circ \ldots \circ \widehat{F}_{d-1}\big)C_i$, and the relation (8.16) becomes

$$\big(\widehat{F}_1 \circ \ldots \circ \widehat{F}_{d-1}\big)\Big(\sum_{i=1}^{d} \dot z_i C_i\Big) = A\big(\psi(z)\big). \tag{8.19}$$

In the situations which we have in mind, the operators $\widehat{F}_j$ can be efficiently inverted, and Algorithm 5.3 can be applied to the solution of (8.1).

The main difficulty of using this coordinate transform is to find a suitable ordering of a basis such that condition (8.18) is satisfied. The following lemma simplifies this task. We use the notation $\alpha_k(C)$ for the coefficient in the representation $C = \sum_{k=1}^{d} \alpha_k(C)C_k$.

Lemma 8.9. *Let* $\{C_1, \ldots, C_d\}$ *be a basis of the Lie algebra* $\mathfrak{g}$. *If for every pair* $j < i$ *and for* $k < j$ *we have*

$$\alpha_k(F_jC_i) \neq 0 \qquad \Longrightarrow \qquad F_\ell C_k = C_k \quad \textit{for } \ell \textit{ satisfying } k \le \ell < j, \tag{8.20}$$

then the relation (8.18) holds for all $i = 2, \ldots, d$.

Proof. We write $\widehat{F}_{i-1}C_i = F_{i-1}C_i = \sum_k \alpha_k(F_{i-1}C_i)C_k$. It follows from the definition of $\widehat{F}_j$ and from (8.20) that $(\widehat{F}_{i-2}\circ\widehat{F}_{i-1})C_i = (F_{i-2}\circ F_{i-1})C_i$. A repeated application of this argument proves the statement. □

Owren & Marthinsen (2001) have studied Lie algebras that admit a basis satisfying (8.18) for all z. We present here one of their examples.

Example 8.10 (Special Linear Group). Consider the differential equation (8.1) on the Lie group $\mathrm{SL}(n) = \{Y \mid \det Y = 1\}$, i.e., the matrix $A(Y)$ lies in $\mathfrak{sl}(n) = \{A \mid \mathrm{trace} A = 0\}$. As a basis of the Lie algebra $\mathfrak{sl}(n)$ we choose $E_{ij} = e_i e_j^T$ for $i \neq j$, and $D_i = e_i e_i^T - e_{i+1}e_{i+1}^T$ for $1 \le i < n$ (here, $e_i = (0, \ldots, 1, \ldots, 0)^T$ denotes the vector whose only non-zero element is in the ith position). Following Owren & Marthinsen (2001) we order the elements of this basis as

$$\begin{aligned} &E_{12} < \ldots < E_{1n} < E_{23} < \ldots < E_{2n} < \ldots < E_{n-1,n} \\ &\quad < E_{21} < \ldots < E_{n1} < E_{32} < \ldots < E_{n2} < \ldots < E_{n,n-1} \\ &\quad\quad < D_1 < \ldots < D_{n-1}. \end{aligned}$$

With the use of Lemma 8.9 one can check in a straightforward way that the relation (8.18) is satisfied. In nearly all situations $\alpha_k(F_jC_i) = 0$ for $k < j < i$, so that (8.18) represents an empty condition. Consequently, the $\dot z_i$ can be computed from (8.19). Due to the sparsity of the matrices E_{ij} and D_i, the computation of $\widehat{F}_i^{-1}$ can be done very efficiently.

IV.9 Geometric Numerical Integration Meets Geometric Numerical Linear Algebra

The persistent use of orthogonal transformations is a hallmark of numerical linear algebra. Correspondingly, manifolds incorporating orthogonality constraints play an important role all over this field; see Edelman, Arias & Smith (1998) on the geometry of algorithms with orthogonality constraints. In addition to the orthogonal group $\mathrm{O}(n)$, the manifolds of primary interest are:

$\mathcal{V}_{n,k}$, the Stiefel manifold of $n \times k$ matrices with k orthonormal columns,
$\mathcal{G}_{n,k}$, the Grassmann manifold of orthogonal projections of $\mathbb{R}^n$ onto k-dimensional subspaces, and
$\mathcal{M}_k^{m\times n}$, the manifold of $m \times n$ matrices of rank k, which is related to orthogonal transformations via the singular value decomposition and a related decomposition discussed below.

IV.9.1 Numerical Integration on the Stiefel Manifold

The original motivation for Stiefel Manifolds (in Stiefel 1935) was the topological problem, whether a manifold $\mathcal{M}$ can possess k everywhere linearly independent continuous vector fields. The problem, which had been solved for the case $k = 1$, was much harder for $k > 1$. In order to attack this question, Stiefel introduced 'his' manifold

$$\mathcal{V}_{n,k} = \{Y \in \mathbb{R}^{n\times k} \,|\, Y^T Y = I\}, \tag{9.1}$$

as an auxiliary tool for the definition of what later became known as the Stiefel-Whitney classes[6].



Eduard Stiefel[5]

Here, we are interested in computations on these manifolds for their own, with many applications, as for example the computation of Lyapunov exponents of differential equations; see Exercise 22 as well as Bridges & Reich (2001) and Dieci, Russell & van Vleck (1997). There are also many cases where orthogonality constraints concern only some of the variables in a differential equation. In molecular dynamics, for example, such orthogonality constraints arise in the Car-Parrinello approach to *ab initio* molecular dynamics (Car & Parrinello 1985) and in the multiconfiguration time-dependent Hartree method of quantum molecular dynamics (Beck, Jäckle, Worth & Meyer 2000).

[5] Eduard L. Stiefel, born: 21 April 1909 in Zürich, died: 25 November 1979; photo: Bildarchiv ETH-Bibliothek, Zürich.
[6] We are grateful to our colleague A. Haefliger for this indication.

Tangent and Normal Space. We choose a fixed matrix Y in the Stiefel manifold $\mathcal{V} = \mathcal{V}_{n,k}$. Then the tangent space (5.4) at $Y \in \mathcal{V}$ consists of the matrices Z such that $(Y + \varepsilon Z)^T (Y + \varepsilon Z)$ remains I for $\varepsilon \to 0$. Differentiating we obtain

$$T_Y \mathcal{V} = \{Z \in \mathbb{R}^{n \times k} \mid Z^T Y + Y^T Z = 0\}, \tag{9.2}$$

i.e., $Y^T Z$ is skew-symmetric. This represents $\frac{1}{2}k(k+1)$ conditions, thus $T_Y \mathcal{V}$ is of dimension $nk - \frac{1}{2}k(k+1)$.

For defining the normal space, we use the standard Euclidean inner product on $\mathbb{R}^{n \times k}$, i.e.,

$$\langle A, B \rangle = \text{trace}(A^T B) = \textstyle\sum_{ij} a_{ij} b_{ij}, \tag{9.3}$$

whose corresponding norm is the Frobenius norm

$$\|A\|_F = \sqrt{\textstyle\sum_{ij} a_{ij}^2}. \tag{9.4}$$

Then the normal space at Y is given by

$$N_Y \mathcal{V} = \{K \in \mathbb{R}^{n \times k} \mid K \perp T_Y \mathcal{V}\} = \{YS \mid S \text{ symmetric } k \times k \text{ matrix}\}. \tag{9.5}$$

To show this, we observe that the orthogonality $YS \perp T_Y \mathcal{V}$ follows from $\langle YS, Z \rangle = \text{trace}(SY^T Z) = \langle S, Y^T Z \rangle$ and the fact that any symmetric matrix A is orthogonal to any skew-symmetric matrix B.[7] A dimension count (the matrix S has $\frac{1}{2}k(k+1)$ free elements) now shows us that the space defined in (9.5) fills the entire orthogonal complement of $T_Y \mathcal{V}$.

Orthogonality-Preserving Runge–Kutta Methods. Suppose now that we have to solve a differential equation $\dot{Y} = F(Y)$ on a Stiefel manifold $\mathcal{V}$. The orthogonality constraints $Y^T Y = I$ are preserved, if the derivative $F(Y)$ lies in the tangent space $T_Y \mathcal{V}$, i.e., if $F(Y)^T Y + Y^T F(Y) = 0$, for every $Y \in \mathcal{V}$ (weak invariants, see Sect. IV.4). In the (exceptional) case where they are in fact true invariants, i.e., if $F(Y)^T Y + Y^T F(Y) = 0$ for *all* $Y \in \mathbb{R}^{n \times k}$, then the orthogonality constraints are quadratic, and are therefore preserved exactly by the implicit Runge–Kutta methods of Sect. IV.2.1, in particular the Gauss methods. These methods give numerical solutions on the Stiefel manifold, but use function evaluations outside the manifold.

In the general case of only weak invariants, a standard approach for enforcing orthogonality is the introduction of *Lagrange multipliers*, which can be interpreted as artificial forces in the direction of the normal space keeping the solutions on the manifold. Due to the structure of $N_Y \mathcal{V}$ (see (9.5)), the problem becomes here

$$\dot{Y} = F(Y) + Y\Lambda, \qquad Y^T Y = I \tag{9.6}$$

with a symmetric Lagrange multiplier matrix $\Lambda \in \mathbb{R}^{k \times k}$; see also Exercise 10. Any numerical method for differential-algebraic equations can now be applied, e.g.,

[7] Indeed, split the sum in (9.3) in two parts $i < j$ and $i > j$, and interchange $i \leftrightarrow j$ in the second sum. Then both sums are identical with opposite sign.

appropriate Runge-Kutta methods as in Chap. VI and Sect. VII.4 of Hairer & Wanner (1996). A symmetric adaptation of Gauss methods to such problems is given by Jay (2005).

Below we shall study in great detail mechanical systems with constraints (see Sect. VII.1). In the case of orthogonality constraints, such problems can be treated successfully with Lobatto IIIA-IIIB partitioned Runge–Kutta methods, which in addition to orthogonality preserve other important geometric properties such as reversibility and symplecticity.

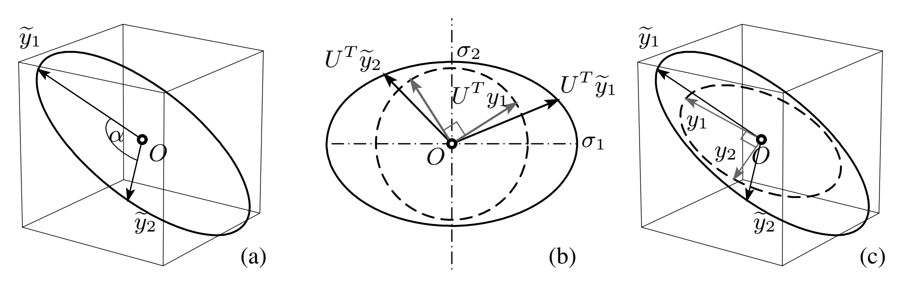


Fig. 9.1. Projection onto the Stiefel manifold using the singular value decomposition

Projection Methods. If we want to use the projection method of Algorithm 4.2, we have to perform, after every integration step, the projection (4.4), which requires to find for any given matrix $\widetilde{Y}$ a matrix $Y \in \mathcal{V}$ with

$$\|Y - \widetilde{Y}\|_F \to \min . \tag{9.7}$$

This projection can be obtained as follows: if $\widetilde{Y}$ is not in $\mathcal{V}$ (but close), then its column vectors $\widetilde{y}_1, \ldots, \widetilde{y}_k$ will have norms different from 1 and/or their angles will not be right angles. These quantities determine an ellipsoid, if we require that these vectors represent conjugate diameters [8] (see Fig. 9.1 (a)). This ellipsoid is then transformed to principal axes in $\mathbb{R}^k$ by an orthogonal map U^T (picture (b)). We let $\sigma_1, \ldots, \sigma_k$ be the length of these axes. If the coordinates are now divided by σ_i, then the ellipsoid becomes the unit sphere and the vectors $U^T\widetilde{y}_i$ become orthonormal vectors $U^T y_i$. These vectors, when transformed back with U, lie in $\mathcal{V}$ and are the projection we were searching for (picture (c)). For a proof of the optimality, see Exercise 21.

Connection with the Singular Value Decomposition. We have by construction that $U^T y_i = \Sigma^{-1} U^T \widetilde{y}_i$ where $\Sigma = \mathrm{diag}(\sigma_1, \ldots \sigma_k)$. If we finally map these vectors by an orthogonal matrix V to the unit base, we see that $V\Sigma^{-1}U^T\widetilde{Y} = I$, or

$$\widetilde{Y} = U\Sigma V^T \tag{9.8}$$

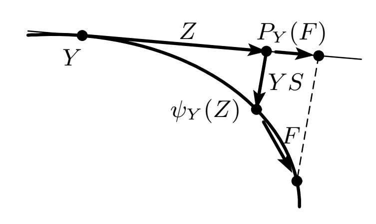
which is the *singular value decomposition* of $\widetilde{Y}$. This connection allows us to use standard software for our calculations. The projected matrix is then $Y = UV^T$.

[8] Here we touch another of Stiefel's great ideas, the CG algorithm.

Remark 1. When the differential equation possesses some symmetry (see the next chapter), then the *symmetric* projection algorithm V.4.1 is preferable to be used instead.

Remark 2. The above procedure is equivalent to the one proposed by D. Higham (1997): the orthogonal projection is the first factor of the *polar decomposition* $\widetilde{Y} = YR$ (where Y has orthonormal columns and R is symmetric positive definite). The equivalence is seen from the polar decomposition $\widetilde{Y} = (UV^T)(V\Sigma V^T)$. A related procedure, where the first factor of the QR decomposition of $\widetilde{Y}$ is used instead of that of the polar decomposition, is proposed in Dieci, Russell & van Vleck (1994).

Tangent Space Parametrization. For the application of the methods of Sect. IV.5, in particular Subsection IV.5.3, to the case of Stiefel manifolds, we have to find the formulas for the projection (5.8) (see the wrap figure).



For a fixed Y, let $Y+Z$ be an arbitrary matrix in $Y + T_Y\mathcal{V}$, for which we search the projection $\psi_Y(Z)$ to $\mathcal{V}$. Because of the structure of $N_Y\mathcal{V}$ (see (9.5)), we have that

$$\psi_Y(Z) = Y + Z + YS \tag{9.9}$$

is a local parametrization of $\mathcal{V}$, if S is symmetric and if $\psi_Y(Z)^T\psi_Y(Z) = I$. This condition, when multiplied out, shows that S has to be a solution of the algebraic Riccati equation

$$S^2 + 2S + SY^TZ + Z^TYS + Z^TZ = 0. \tag{9.10}$$

Observe that for $k = 1$, where the Stiefel manifold reduces to the unit sphere in $\mathbb{R}^n$, the equation (9.10) is a scalar quadratic equation and can be easily solved. For $k > 1$, it can be solved iteratively using the scheme (e.g., starting with $S_0 = 0$)

$$(I + Z^TY)S_n + S_n(I + Y^TZ) = -Z^TZ - S_{n-1}^2.$$

Using a Schur decomposition $Y^TZ = Q^TRQ$ (where Q is orthogonal and R upper triangular), the elements of QS_nQ^T can be computed successively starting from the left upper corner. We refer to the monograph of Mehrmann (1991) for a detailed discussion of the solution of linear and algebraic Riccati equations.

Next, we compute for the matrix F its orthogonal projection $P_Y(F)$ to $T_Y\mathcal{V}$, i.e., by (9.5), we have to find a symmetric matrix $\widetilde{S}$ such that $P_Y(F) = F - Y\widetilde{S}$. The tangent condition $P_Y(F)^TY + Y^TP_Y(F) = 0$ leads to $\widetilde{S} = (F^TY + Y^TF)/2$, so that

$$P_Y(F) = F - \frac{1}{2}\big(YF^TY + YY^TF\big). \tag{9.11}$$

With the parametrization $\psi_Y(Z)$ of (9.9) the transformed differential equation, when projected to the tangent space, yields

$$\dot{Z} = P_YF\big(\psi_Y(Z)\big), \tag{9.12}$$

in complete analogy to (5.9). The numerical solution of (9.12) requires, for every function evaluation, the solution of the Riccati equation (9.10) and the computation of a projection onto the tangent space, each needing $\mathcal{O}(nk^2)$ operations. Compared with the projection method, the overhead (i.e., the computation apart from the evaluation of $F(Y)$) is more expensive, but the approach described here has the advantage that all evaluations of F are exactly on the manifold $\mathcal{V}$.

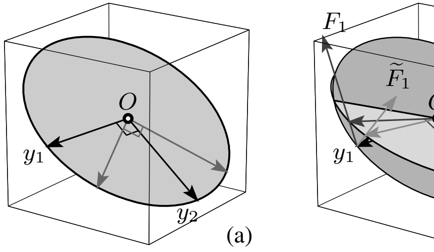
IV.9.2 Differential Equations on the Grassmann Manifold

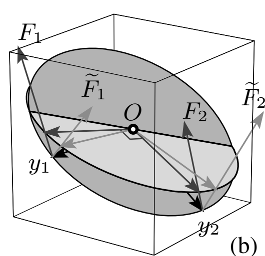
The Grassmann manifold is obtained from the Stiefel manifold by identifying matrices in $\mathcal{V}_{n,k}$ that span the same subspace (see Fig. 9.2 (a)). Since any two such matrices result from each other by right multiplication with an orthogonal $k \times k$ matrix, the resulting manifold is the quotient manifold

$$\mathcal{G}_{n,k} = \mathcal{V}_{n,k}/\mathrm{O}(k). \tag{9.13}$$

An equivalence class $[Y] \in \mathcal{G}_{n,k}$ defines an orthogonal projection $P = YY^T$ of rank k, and conversely, every orthogonal basis of the range of P yields a representative $Y \in \mathcal{V}_{n,k}$. We can thus view the Grassmann manifold as

$$\mathcal{G}_{n,k} = \left\{ P \;\middle|\; \begin{array}{l} P \text{ is an orthogonal projection onto} \\ \text{a } k\text{-dimensional subspace of } \mathbb{R}^n \end{array} \right\}. \tag{9.14}$$





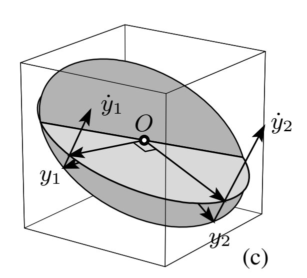


Fig. 9.2. Integration of a differential equation on the Grassmann manifold

The Tangent Space. The map $Y \mapsto P = YY^T$ from $\mathcal{V} \to \mathcal{G}$ has the tangent map (derivative) [9]

$$T_Y\mathcal{V} \to T_P\mathcal{G}: \quad \delta Y \mapsto \delta P = \delta Y Y^T + Y \delta Y^T, \tag{9.15}$$

and we wish to apply all the methods for $T_Y\mathcal{V}$ from the arsenal of the preceding section to problems in $T_P\mathcal{G}$. However, the dimension of $T_P\mathcal{G}$ is by $\frac{1}{2}k(k-1)$ lower than the dimension of $T_Y\mathcal{V}$. This difference is the dimension of $\mathrm{O}(k)$ and also of

[9] Here we write δY for tangent matrices at Y (what has been Z in (9.2)), and similarly for other matrices; Lagrange's δ-notation here becomes preferable, since we will have, especially in the next subsection, more and more matrices moving around.

$\mathfrak{so}(k)$, the vector space of skew-symmetric $k \times k$ matrices. The key idea is now the following: if we replace the condition from (9.2), $Y^T \delta Y$ skew-symmetric, by $Y^T \delta Y = 0$, then we remove precisely the superfluous degrees of freedom. Indeed, the extended tangent map

$$T_Y \mathcal{V} \to T_P \mathcal{G} \times \mathfrak{so}(k) : \quad \delta Y \mapsto (\delta Y Y^T + Y \delta Y^T, Y^T \delta Y) \tag{9.16}$$

is an isomorphism, since it is readily seen to have zero null-space and the dimensions of the vector spaces agree. The tangent space is thus characterized as

$$T_P \mathcal{G} = \{\delta P = \delta Y Y^T + Y \delta Y^T \mid Y^T \delta Y = 0\}, \tag{9.17}$$

and every $\delta P \in T_P \mathcal{G}$ corresponds to a unique δY with $Y^T \delta Y = 0$. Note that this condition on δY does not depend on the representative Y of $[Y]$.

Differential Equations. Consider now a differential equation on $\mathcal{G}$,

$$\dot{P} = G(P), \tag{9.18}$$

with a vector field G on $\mathcal{G}$. The condition $G(P) \in T_P \mathcal{G}$ means, since the tangent map (9.15) is onto, that there exists for $P = YY^T$ a vector $F(Y)$ such that

$$G(P) = F(Y)Y^T + YF(Y)^T \qquad \text{with} \qquad F^T Y + Y^T F = 0 \tag{9.19}$$

i.e., $F(Y) \in T_Y \mathcal{V}$. However, from a given initial position Y, there are many F which produce the same movement G of the subspace represented by P (see Fig. 9.2 (b)). By (9.16), the movement of Y becomes unique if we require that this movement is *orthogonal* to the subspace (see Fig. 9.2 (c)),

$$Y^T \dot{Y} = 0\,. \tag{9.20}$$

Multiplying the derivative $\dot{P} = \dot{Y} Y^T + Y \dot{Y}^T$ with Y^T from the left, we obtain, under condition (9.20), $Y^T \dot{P} = \dot{Y}^T$ and, by (9.18) and (9.19), $\dot{Y} = Y F^T Y + F$ or

$$\dot{Y} = (I - YY^T) F(Y). \tag{9.21}$$

Geometrically, this means that the vector $F(Y)$, which could be chosen arbitrarily in $T_Y \mathcal{V}$, is projected to the orthogonal complement of the subspace spanned by Y or $P = YY^T$. The derivative $\dot{Y}$ in (9.21) is independent of the particular choice of F.

Equation (9.21) is a differential equation on the Stiefel manifold $\mathcal{V}$ that can be solved numerically by the methods described in the previous subsection.

Example 9.1 (Oja Flow). A basic example arises in neural networks (Oja 1989): solutions on $\mathcal{V}_{n,k}$ of the differential equation

$$\dot{Y} = (I - YY^T) AY \tag{9.22}$$

with a constant symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ tend to an orthogonal basis of an invariant subspace of A as $t \to \infty$ (Yan, Helmke & Moore 1994).

A naïve comparison of this equation with (9.21) would lead to $F(Y) = AY$, but this function does not satisfy the tangent condition $F^T Y + Y^T F = 0$ from (9.19). So we use the fact that $(I - YY^T)^2 = I - YY^T$ and set $F(Y) = (I - YY^T)AY$. With this, $G(P)$ from (9.18) and (9.19) becomes

$$\dot{P} = (I - P)AP + PA(I - P). \tag{9.23}$$

We have obtained the result that equation (9.22) can be viewed as a differential equation on the Grassmann manifold $\mathcal{G}_{n,k}$.

However, for the numerical integration it is more practical to work with (9.22).

IV.9.3 Dynamical Low-Rank Approximation

Low-rank approximation of large matrices is a basic model reduction technique in many application areas, such as image compression and latent semantic indexing in information retrieval; see for example Simon & Zha (2000). Here, we consider the task of computing low rank approximations to matrices $A(t) \in \mathbb{R}^{m\times n}$ depending smoothly on t. At any time t, a best approximation to $A(t)$ of rank k is a matrix $X(t)$ in the manifold $\mathcal{M}_k = \mathcal{M}_k^{m\times n}$ of rank-k matrices that satisfies

$$X(t) \in \mathcal{M}_k \quad \text{such that} \quad \|X(t) - A(t)\|_F = \min! \tag{9.24}$$

The problem is solved by a singular value decomposition of $A(t)$, truncating all singular values after the r largest ones. When the matrix is so large that a complete singular value decomposition is not feasible, a standard approach to obtain an approximate solution is based on the Lanczos bidiagonalization process with $A(t)$, as discussed in Simon & Zha (2000).

Following Koch & Lubich (2005), we here consider instead the low-rank approximation $Y(t) \in \mathcal{M}_k$ determined from the condition that for every t the derivative $\dot{Y}(t)$, which is in the tangent space $T_{Y(t)}\mathcal{M}_k$, be chosen as

$$\dot{Y}(t) \in T_{Y(t)}\mathcal{M}_k \quad \text{such that} \quad \|\dot{Y}(t) - \dot{A}(t)\|_F = \min! \tag{9.25}$$

This is complemented with an initial condition, ideally $Y(t_0) = X(t_0)$. For given $Y(t)$, the derivative $\dot{Y}(t)$ is obtained by a *linear* projection, though onto a solution-dependent vector space. Problem (9.25) yields a differential equation on $\mathcal{M}_k$. We will see that with a suitable factorization of rank-k matrices, we obtain a system of differential equations for the factors that is well-suited for numerical integration. The differential equations contain only the increments $\dot{A}(t)$, which may be much sparser than the full data matrix $A(t)$.

Koch & Lubich (2005) show that $Y(t)$ yields a quasi-optimal approximation on intervals where a good smooth approximation exists. It must be noted, however, that the best rank-k approximation $X(t)$ may have discontinuities, which cannot be captured in $Y(t)$. This is already seen from the example of finding a rank-1 approximation to $\text{diag}(e^{-t}, e^t)$, where starting from $t_0 < 0$ yields $X(t) = Y(t) = \text{diag}(e^{-t}, 0)$ for $t < 0$, but $Y(t) = \text{diag}(e^{-t}, 0)$ and $X(t) = \text{diag}(0, e^t)$ for $t > 0$.

The best approximation $X(t)$ has a discontinuity at $t = 0$, caused by a crossing of singular values of which one is inside and the other one outside the approximation. An algorithmic remedy is to restart (9.25) at regular intervals.

In contrast to (9.24), the approach (9.25) extends immediately to the low-rank approximation of solutions of matrix differential equations $\dot{A} = F(A)$. Here, $\dot{A}(t)$ in (9.25) is simply replaced by the approximation $F(Y(t))$, which yields the minimum-defect low-rank approximation $Y(t)$ by choosing

$$\dot{Y} \in T_Y\mathcal{M}_k \quad \text{such that} \quad \|\dot{Y} - F(Y)\|_F = \min! \tag{9.26}$$

An approach of this type is of common use in quantum dynamics, where the physical model reduction of the multivariate Schrödinger equation by the analogue of (9.26) is known as the Dirac-Frenkel time-dependent variational principle, after Dirac (1930) and Frenkel (1934); see also Beck, Jäckle, Worth & Meyer (2000) and Sect. VII.6.

Decompositions of Rank-k Matrices and of Their Tangent Matrices. Every real rank-k matrix of dimension $m \times n$ can be written in the form

$$Y = USV^T \tag{9.27}$$

where $U \in \mathcal{V}_{m,k}$ and $V \in \mathcal{V}_{n,k}$ have orthonormal columns, and $S \in \mathbb{R}^{k\times k}$ is nonsingular. The singular value decomposition yields S diagonal, but here we do not assume a special form of S. The representation (9.27) is not unique: replacing U by $\widetilde{U} = UP$ and V by $\widetilde{V} = VQ$ with orthogonal matrices $P, Q \in \mathrm{O}(k)$ and correspondingly S by $\widetilde{S} = P^TSQ$, yields the same matrix $Y = USV^T = \widetilde{U}\widetilde{S}\widetilde{V}^T$.

As a substitute for the non-uniqueness in (9.27), we use – as in the previous subsection – a unique decomposition in the tangent space. Every tangent matrix $\delta Y \in T_Y\mathcal{M}_k$ at $Y = USV^T$ is of the form (see Exercise 23)

$$\delta Y = \delta USV^T + U\delta SV^T + US\delta V^T, \tag{9.28}$$

where $\delta S \in \mathbb{R}^{k\times k}$ and $\delta U \in T_U\mathcal{V}_{m,k}$, $\delta V \in T_V\mathcal{V}_{n,k}$. Conversely, $\delta S, \delta U, \delta V$ are uniquely determined by δY if we impose the orthogonality constraints

$$U^T\delta U = 0, \qquad V^T\delta V = 0. \tag{9.29}$$

Equations (9.28) and (9.29) yield

$$\begin{aligned} \delta S &= U^T\delta YV, \\ \delta U &= (I - UU^T)\delta YVS^{-1}, \\ \delta V &= (I - VV^T)\delta Y^TUS^{-T}. \end{aligned} \tag{9.30}$$

Formulas (9.28) and (9.30) establish an isomorphism between the subspace

$$\{(\delta S, \delta U, \delta V) \in \mathbb{R}^{k\times k} \times \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} \,|\, U^T\delta U = 0,\ V^T\delta V = 0\}$$

and the tangent space $T_Y\mathcal{M}_k$.

Differential Equations for the Factors. The minimization condition (9.25) is equivalent to the orthogonal projection of $\dot{A}(t)$ onto the tangent space $T_{Y(t)}\mathcal{M}_k$: find $\dot{Y} \in T_Y\mathcal{M}_k$ (we omit the argument t) satisfying

$$\langle \dot{Y} - \dot{A},\, \delta Y \rangle = 0 \quad \text{for all } \ \delta Y \in T_Y\mathcal{M}_k, \tag{9.31}$$

with the Frobenius inner product $\langle A, B\rangle = \text{trace}(A^T B)$. With this formulation we derive differential equations for the factors in the representation (9.27).

Theorem 9.2. *For $Y = USV^T \in \mathcal{M}_k$ with nonsingular $S \in \mathbb{R}^{k\times k}$ and with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ having orthonormal columns, condition (9.25) or (9.31) is equivalent to $\dot{Y} = \dot{U}SV^T + U\dot{S}V^T + US\dot{V}^T$, where*

$$\begin{aligned} \dot{S} &= U^T \dot{A} V \\ \dot{U} &= (I - UU^T)\dot{A}VS^{-1} \\ \dot{V} &= (I - VV^T)\dot{A}^T U S^{-T}. \end{aligned} \tag{9.32}$$

Proof. For $u \in \mathbb{R}^m,\ v \in \mathbb{R}^n$ and $B \in \mathbb{R}^{m\times n}$, we use the identity

$$\langle uv^T, B\rangle = u^T B v.$$

In view of (9.29) we require $U^T\dot{U} = V^T\dot{V} = 0$ along the solution trajectory in order to define a unique representation of $\dot{Y}$. We first substitute $\delta Y = u_i v_j^T$, for $i,\ j = 1,\ldots,k$, in (9.31), where u_i, v_j denote the columns of $U,\ V$, respectively. This is of the form (9.27) with $\delta U = \delta V = 0$ and one non-zero element in δS. In this way we find $\dot{S} = U^T\dot{A}V$. Similarly, choosing $\delta Y = \sum_{j=1}^k \delta u\, s_{ij} v_j^T,\ i = 1,\ldots,k$, where $\delta u \in \mathbb{R}^m$ is arbitrary with $U^T\delta u = 0$, we obtain the stated differential equation for U, and likewise for $\delta Y = \sum_{j=1}^k u_j s_{ji} \delta v^T$ with $V^T \delta v = 0$ the differential equation for V. □

The differential equations (9.32) are closely related to differential equations for other smooth matrix decompositions, in particular the smooth singular value decomposition; see, e.g., Dieci & Eirola (1999) and Wright (1992). Unlike the differential equations for singular values given there, the equations (9.32) have no singularities at points where singular values of $Y(t)$ coalesce.

For the minimum-defect low-rank approximation (9.26) of a matrix differential equation $\dot{A} = F(A)$, we just need to replace $\dot{A}$ by $F(Y)$ for $Y = USV^T$ in the differential equations (9.32).

The matrices $U(t)$ and $V(t)$ evolve on Stiefel manifolds. The differential equations (9.32) can thus be solved numerically by the methods discussed in Sect. IV.9.1.

IV.10 Exercises

1. Prove that the symplectic Euler method (I.1.9) conserves quadratic invariants of the form (2.5). Explain the “0” entries of Table (I.2.1).

2. Prove that under condition (2.3) a Runge–Kutta method preserves all invariants of the form $I(y) = y^TCy + d^Ty + c$.
3. Prove that an s-stage diagonally implicit Runge–Kutta method (i.e., $a_{ij} = 0$ for $i < j$) satisfies the condition (2.3) if and only if it is equivalent to a composition $\Phi_{b_s h} \circ \ldots \circ \Phi_{b_1 h}$ based on the implicit midpoint rule.
4. Prove the following statements: a) If a partitioned Runge–Kutta method conserves general quadratic invariants $p^TCp + 2p^TDq + q^TEq$, then each of the two Runge–Kutta methods has to conserve quadratic invariants separately.
   b) If both methods, $\{b_i, a_{ij}\}$ and $\{\widehat{b}_i, \widehat{a}_{ij}\}$ are irreducible, satisfy (2.3) and if (2.7)-(2.8) hold, then we have $b_i = \widehat{b}_i$ and $a_{ij} = \widehat{a}_{ij}$ for all i, j.
5. Prove that the Gauss methods are the only collocation methods satisfying (2.3).
   Hint. Use the ideas of the proof of Lemma 13.9 in Hairer & Wanner (1996).
6. Discontinuous collocation methods with either $b_1 \neq 0$ or $b_s \neq 0$ (Definition II.1.7) cannot satisfy the criterion (2.3).
7. (Sanz-Serna & Abia 1991, Saito, Sugiura & Mitsui 1992). The condition (2.3) acts as simplifying assumption for the order conditions of Runge–Kutta methods. Assume that the order conditions are satisfied for the trees u and v. Prove that it is satisfied for $u \circ v$ if and only if it is satisfied for $v \circ u$, and that it is automatically satisfied for trees of the form $u \circ u$.
   Remark. $u \circ v$ denotes the Butcher product introduced in Sect. VI.7.2.
8. If L_0 is a symmetric, tridiagonal matrix that is sufficiently close to $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ are the eigenvalues of L_0, then the solution of (3.5) with $B(L) = L_+ - L_+^T$ converges exponentially fast to the diagonal matrix Λ. Hence, the numerical solution of (3.5) gives an algorithm for the computation of the eigenvalues of the matrix L_0.
   Hint. Let $\beta_1, \ldots, \beta_n$ be the entries in the diagonal of L, and $\alpha_1, \ldots, \alpha_{n-1}$ those in the subdiagonal. Assume that $|\beta_k(0) - \lambda_k| \leq R/3$ and $|\alpha_k(0)| \leq R$ with some sufficiently small R. Prove that $\beta_k(t) - \beta_{k+1}(t) \geq \mu - R$ and $|\alpha_k(t)| \leq Re^{-(\mu-R)t}$ for all $t \geq 0$, where $\mu = \min_k(\lambda_k - \lambda_{k+1}) > 0$.
9. Elaborate Example 4.5 for the special case where Y is a matrix of dimension 2. In particular, show that (4.6) is the same as (4.5), and check the formulas for the simplified Newton iterations.
10. (Brenan, Campbell & Petzold (1996), Sect. 2.5.3). Consider the differential equation $\dot{y} = f(y)$ with known invariants $g(y) = Const$, and assume that $g'(y)$ has full rank. Prove by differentiation of the constraints that, for initial values satisfying $g(y_0) = 0$, the solution of the differential-algebraic equation (DAE)

$$\begin{aligned} \dot{y} &= f(y) + g'(y)^T\mu \\ 0 &= g(y) \end{aligned}$$

    also solves the differential equation $\dot{y} = f(y)$.
    Remark. Most methods for DAEs (e.g., stiffly accurate Runge–Kutta methods or BDF methods) lead to numerical integrators that preserve exactly the constraints $g(y) = 0$. The difference from the projection method of Sect. IV.4 is that here the internal stages also satisfy the constraint.

11. Prove that $\mathrm{SL}(n)$ is a Lie group of dimension n^2-1, and that $\mathfrak{sl}(n)$ is its Lie algebra (see Table 6.1 for the definitions of $\mathrm{SL}(n)$ and $\mathfrak{sl}(n)$).
12. Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. Prove that for $Y \in G$ and $A \in \mathfrak{g}$ we have $YAY^{-1} \in \mathfrak{g}$.
    Hint. Consider the path $\gamma(t) = Y\alpha(t)Y^{-1}$.
13. Consider a problem $\dot{Y} = A(Y)Y$, for which $A(Y) \in \mathfrak{so}(n)$ whenever $Y \in \mathrm{O}(n)$, but where $A(Y)$ is an arbitrary matrix for $Y \notin \mathrm{O}(n)$.
    a) Prove that $Y_0 \in \mathrm{O}(n)$ implies $Y(t) \in \mathrm{O}(n)$ for all t.
    b) Show by a counter-example that the numerical solution of the implicit midpoint rule does not necessarily stay in $\mathrm{O}(n)$.
14. (Feng Kang & Shang Zai-jiu 1995). Let $R(z) = (1+z/2)/(1-z/2)$ be the stability function of the implicit midpoint rule. Prove that for $A \in \mathfrak{sl}(3)$ we have
$$\det R(hA) = 1 \quad \Leftrightarrow \quad \det A = 0.$$
15. (Iserles & Nørsett 1999). Introducing $y_1 = y$ and $y_2 = \dot{y}$, write the problem
$$\ddot{y} + ty = 0, \qquad y(0) = 1, \qquad \dot{y}(0) = 0$$
    in the form (7.6). Then apply the numerical method of Example 7.4 with different step sizes on the interval $0 \le t \le 100$. Compare the result with that obtained by fourth order classical (explicit or implicit) Runge–Kutta methods.
    Remark. If $A(t)$ in (7.6) (or $A(t,y)$ in (8.1)) are much smoother than the solution $y(t)$, then Lie group methods are usually superior to standard integrators, because Lie group methods approximate $A(t)$, whereas standard methods approximate the solution $y(t)$ by polynomials.
16. Deduce the BCH formula from the Magnus expansion (IV.7.5).
    Hint. For constant matrices A and B consider the matrix function $A(t)$ defined by $A(t) = B$ for $0 \le t \le 1$ and $A(t) = A$ for $1 \le t \le 2$.
17. (Rodrigues formula, see Marsden & Ratiu (1999), page 291). Prove that
$$\exp(\Omega) = I + \frac{\sin\alpha}{\alpha}\,\Omega + \frac{1}{2}\Big(\frac{\sin(\alpha/2)}{\alpha/2}\Big)^2\Omega^2 \quad \text{for} \quad \Omega = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$$
    where $\alpha = \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2}$. This formula allows for an efficient implementation of the Lie group methods in $\mathrm{O}(3)$.
18. The solution of $\dot{Y} = A(Y)Y, Y(0) = Y_0$, is given by $Y(t) = \exp\big(\Omega(t)\big)Y_0$, where $\Omega(t)$ solves the differential equation (8.9). Compute the first terms of the t-expansion of $\Omega(t)$.
    Result. $\Omega(t) = tA(Y_0) + \frac{t^2}{2}A'(Y_0)A(Y_0)Y_0 + \frac{t^3}{6}\big(A'(Y_0)^2A(Y_0)Y_0^2 + A'(Y_0)A(Y_0)^2Y_0 + A''(Y_0)\big(A(Y_0)Y_0, A(Y_0)Y_0\big) - \frac{1}{2}\big[A(Y_0), A'(Y_0)A(Y_0)Y_0\big]\big)$.
19. Consider the 2-stage Gauss method of order $p = 4$. In the corresponding Lie group method, eliminate the presence of Ω in $[\Omega, A]$ by iteration, and neglect higher order commutators. Show that this leads to

$$\begin{aligned}
\Omega_1 &= h\Big(\frac{1}{4}A_1 + \Big(\frac{1}{4} - \frac{\sqrt{3}}{6}\Big)A_2\Big) - \frac{h^2}{2}\Big(-\frac{1}{12} + \frac{\sqrt{3}}{24}\Big)[A_1, A_2] \\
\Omega_2 &= h\Big(\Big(\frac{1}{4} + \frac{\sqrt{3}}{6}\Big)A_1 + \frac{1}{4}A_2\Big) - \frac{h^2}{2}\Big(\frac{1}{12} + \frac{\sqrt{3}}{24}\Big)[A_1, A_2] \\
y_1 &= \exp\Big(h\Big(\frac{1}{2}\,A_1 + \frac{1}{2}\,A_2\Big) - h^2\,\frac{\sqrt{3}}{12}\,[A_1, A_2]\Big)y_0,
\end{aligned}$$

where $A_i = A(Y_i)$ and $Y_i = \exp(\Omega_i)y_0$. Prove that this is a Lie group method of order 4. Is it symmetric?

20. In Zanna (1999) a Lie group method similar to that of Exercise 19 is presented. The only difference is that the coefficients $(-1/12 + \sqrt{3}/24)$ and $(1/12 + \sqrt{3}/24)$ in the formulas for Ω_1 and Ω_2 are replaced with $(-5/72+\sqrt{3}/24)$ and $(5/72 + \sqrt{3}/24)$, respectively. Is there an error somewhere? Are both methods of order 4?
21. Show that for given $\widetilde{Y}$ the solution of problem (9.7) is $Y = UV^T$, where $\widetilde{Y} = U\Sigma V^T$ is the singular value decomposition of $\widetilde{Y}$.
 Hint. Since $\|USV^T\|_F = \|S\|_F$ holds for all orthogonal matrices U and V, it is sufficient to consider the case $\widetilde{Y} = (\Sigma, 0)^T$ with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$. Prove that $\|(\Sigma, 0)^T - Y\|_F^2 \geq \sum_{i=1}^k (\sigma_i - 1)^2$ for all matrices Y satisfying $Y^TY = I$.
22. Show that the solution of the matrix differential equation $\dot{Y} = A(t)Y$ on $\mathbb{R}^{n\times k}$, with initial values $Y_0 \in \mathcal{V}_{n,k}$, can be decomposed as
 $$Y(t) = U(t)S(t), \qquad \text{where } U(t) \in \mathcal{V}_{n,k},\ S(t) \in \mathbb{R}^{k\times k}$$
 satisfy the differential equations
 $$\dot{S} = U^TAUS, \qquad \dot{U} = (I - UU^T)AU$$
 with initial values $S_0 = I$, $U_0 = Y_0$.
 Remark: These differential equations can be used for the computation of Lyapunov exponents as an alternative to the differential equations discussed in Bridges & Reich (2001) and Dieci, Russell & van Vleck (1997).
23. Consider the map $\mathrm{GL}(k) \times \mathcal{V}_{m,k} \times \mathcal{V}_{n,k} \to \mathcal{M}_k$ that associates to (S, U, V) the rank-k matrix $Y = USV^T$. Show that the extended tangent map
 $$\begin{aligned}
 &\mathbb{R}^{k\times k} \times T_U\mathcal{V}_{m,k} \times T_V\mathcal{V}_{n,k} \to T_Y\mathcal{M}_k \times \mathfrak{so}(k) \times \mathfrak{so}(k) \\
 &(\delta S, \delta U, \delta V) \mapsto (\delta USV^T + U\delta SV^T + US\delta V^T, U^T\delta U, V^T\delta V)
 \end{aligned}$$
 is an isomorphism.
24. Let $A(t) \in \mathbb{R}^{n\times n}$ be symmetric and depend smoothly on t. Show that the solution $P(t) \in \mathcal{G}_{n,k}$ of the dynamical low-rank approximation problem on the Grassmann manifold,
 $$\dot{P} \in T_P\mathcal{G}_{n,k} \qquad \text{with} \qquad \|\dot{P} - \dot{A}\|_F = \min!,$$
 is given as $P = YY^T$ where $Y \in \mathcal{V}_{n,k}$ solves the differential equation
 $$\dot{Y} = (I - YY^T)\dot{A}Y.$$

Chapter V. Symmetric Integration and Reversibility

Symmetric methods of this chapter and symplectic methods of the next chapter play a central role in the geometric integration of differential equations. We discuss reversible differential equations and reversible maps, and we explain how symmetric integrators are related to them. We study symmetric Runge–Kutta and composition methods, and we show how standard approaches for solving differential equations on manifolds can be symmetrized. A theoretical explanation of the excellent long-time behaviour of symmetric methods applied to reversible differential equations will be given in Chap. XI.

V.1 Reversible Differential Equations and Maps

Conservative mechanical systems have the property that inverting the initial direction of the velocity vector and keeping the initial position does not change the solution trajectory, it only inverts the direction of motion. Such systems are "reversible". We extend this notion to more general situations.

Definition 1.1. Let ρ be an invertible linear transformation in the phase space of $\dot{y} = f(y)$. This differential equation and the vector field $f(y)$ are called *ρ-reversible* if

$$\rho f(y) = -f(\rho y) \qquad \text{for all } y. \tag{1.1}$$

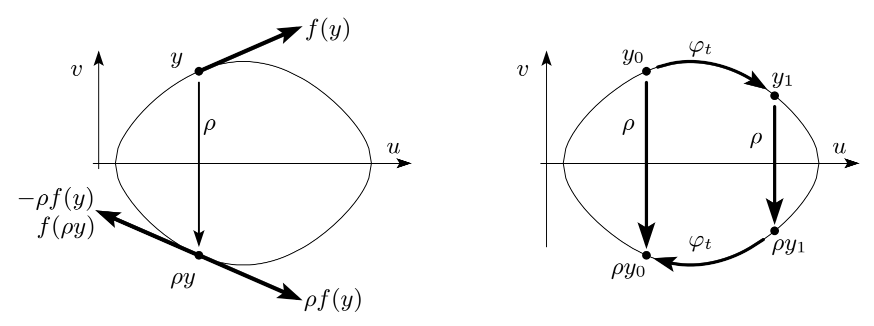


Fig. 1.1. Reversible vector field (left picture) and reversible map (right picture)

This property is illustrated in the left picture of Fig. 1.1. For ρ-reversible differential equations the exact flow $\varphi_t(y)$ satisfies

$$\rho \circ \varphi_t = \varphi_{-t} \circ \rho = \varphi_t^{-1} \circ \rho \tag{1.2}$$

(see the picture to the right in Fig. 1.1). The right identity is a consequence of the group property $\varphi_t \circ \varphi_s = \varphi_{t+s}$, and the left identity follows from

$$\begin{aligned} \frac{d}{dt}\big(\rho \circ \varphi_t\big)(y) &= \rho f\Big(\varphi_t(y)\Big) = -f\Big(\big(\rho \circ \varphi_t\big)(y)\Big) \\ \frac{d}{dt}\big(\varphi_{-t} \circ \rho\big)(y) &= -f\Big(\big(\varphi_{-t} \circ \rho\big)(y)\Big), \end{aligned}$$

because all expressions of (1.2) satisfy the same differential equation with the same initial value $(\rho \circ \varphi_0)(y) = (\varphi_0 \circ \rho)(y) = \rho y$. Formula (1.2) motivates the following definition.

Definition 1.2. A map $\Phi(y)$ is called *ρ-reversible* if

$$\rho \circ \Phi = \Phi^{-1} \circ \rho.$$

Example 1.3. An important example is the partitioned system

$$\dot{u} = f(u,v), \qquad \dot{v} = g(u,v), \tag{1.3}$$

where $f(u,-v) = -f(u,v)$ and $g(u,-v) = g(u,v)$. Here, the transformation ρ is given by $\rho(u,v) = (u,-v)$. If we call a vector field or a map *reversible* (without specifying the transformation ρ), we mean that it is ρ-reversible with this particular ρ. All second order differential equations $\ddot{u} = g(u)$ written as $\dot{u} = v$, $\dot{v} = g(u)$ are reversible. As a first implication of reversibility on the dynamics we mention the following fact: if u and v are scalar, and if (1.3) is reversible, then any solution that crosses the u-axis twice is periodic (Exercise 5, see also the solution of the pendulum problem in Fig. I.1.4).

It is natural to search for numerical methods that produce a reversible numerical flow when they are applied to a reversible differential equation. We then expect the numerical solution to have long-time behaviour similar to that of the exact solution; see Chap. XI for more precise statements. It turns out that the ρ-reversibility of a numerical one-step method is closely related to the concept of symmetry.

> Thus the method is theoretically *symmetrical* or *reversible*, a terminology we have never seen applied elsewhere.
>
> (P.C. Hammer & J.W. Hollingsworth 1955)

Definition 1.4. A numerical one-step method Φ_h is called *symmetric* or *time-reversible*,[1] if it satisfies

$$\Phi_h \circ \Phi_{-h} = id \qquad \text{or equivalently} \qquad \Phi_h = \Phi_{-h}^{-1}.$$

---

[1] The study of symmetric methods has its origin in the development of extrapolation methods (Gragg 1965, Stetter 1973), because the global error admits an asymptotic expansion in even powers of h. The notion of time-reversible methods is more common in the Computational Physics literature (Buneman 1967).

With the Definition II.3.1 of the adjoint method (i.e., $\Phi_h^* = \Phi_{-h}^{-1}$), the condition for symmetry reads $\Phi_h = \Phi_h^*$. A method $y_1 = \Phi_h(y_0)$ is symmetric if exchanging $y_0 \leftrightarrow y_1$ and $h \leftrightarrow -h$ leaves the method unaltered. In Chap. I we have already encountered the implicit midpoint rule (I.1.7) and the Störmer–Verlet scheme (I.1.17), both of which are symmetric. Many more symmetric methods will be given in the following sections.

Theorem 1.5. *If a numerical method, applied to a ρ-reversible differential equation, satisfies*

$$\rho \circ \Phi_h = \Phi_{-h} \circ \rho, \tag{1.4}$$

then the numerical flow Φ_h is a ρ-reversible map if and only if Φ_h is a symmetric method.

Proof. As a consequence of (1.4) the numerical flow Φ_h is ρ-reversible if and only if $\Phi_{-h} \circ \rho = \Phi_h^{-1} \circ \rho$. Since ρ is an invertible transformation, this is equivalent to the symmetry of the method Φ_h. □

Similarly, it is also true that a symmetric method is ρ-reversible if and only if the ρ-compatibility condition (1.4) holds.

Compared to the symmetry of the method, condition (1.4) is much less restrictive. It is automatically satisfied by most numerical methods. Let us briefly discuss the validity of (1.4) for different classes of methods.

- *Runge–Kutta methods* (explicit or implicit) satisfy (1.4) without any restriction other than (1.1) on the vector field (Stoffer 1988). Let us illustrate the proof with the explicit Euler method $\Phi_h(y_0) = y_0 + hf(y_0)$:

$$(\rho \circ \Phi_h)(y_0) = \rho y_0 + h\rho f(y_0) = \rho y_0 - hf(\rho y_0) = \Phi_{-h}(\rho y_0).$$

- *Partitioned Runge–Kutta methods* applied to a partitioned system (1.3) satisfy the condition (1.4) if $\rho(u,v) = \big(\rho_1(u), \rho_2(v)\big)$ with invertible ρ_1 and ρ_2. The proof is the same as for Runge–Kutta methods. Notice that the mapping $\rho(u,v) = (u,-v)$ of Example 1.3 is of this special form.
- *Composition methods*. If two methods Φ_h and Ψ_h satisfy (1.4), then so does the adjoint Φ_h^* and the composition $\Phi_h \circ \Psi_h$. Consequently, the composition methods (3.1) and (3.2) below, which compose a basic method Φ_h and its adjoint with different step sizes, have the property (1.4) provided the basic method Φ_h has it.
- *Splitting methods* are based on a splitting $\dot y = f^{[1]}(y) + f^{[2]}(y)$ of the differential equation. If both vector fields, $f^{[1]}(y)$ and $f^{[2]}(y)$, satisfy (1.1), then their exact flows $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ satisfy (1.2). In this situation, the splitting method (II.5.6) has the property (1.4).
- For *differential equations on manifolds* we have to assume that ρ maps $\mathcal{M}$ to $\mathcal{M}$. Otherwise, condition (1.1) does not make sense. For the projection method of Algorithm IV.4.2 with orthogonal projection onto the manifold we have: if the basic method satisfies (1.4) and if ρ is an orthogonal matrix, then it satisfies (1.4) as well. This follows from the fact that the tangent and normal spaces satisfy

$T_{\rho y}\mathcal{M} = \rho T_y\mathcal{M}$ and $N_{\rho y}\mathcal{M} = \rho^{-T}N_y\mathcal{M}$, respectively. A similar result holds for methods based on local coordinates, if the local parametrization is well chosen. For example, this is the case if $\rho\psi(z)$ is the parametrization at ρy_0 whenever $\psi(z)$ is the parametrization at y_0.

V.2 Symmetric Runge–Kutta Methods

We give a characterization of symmetric methods of Runge–Kutta type and mention some important examples.

V.2.1 Collocation and Runge–Kutta Methods

Symmetric collocation methods are characterized by the symmetry of the collocation points with respect to the midpoint of the integration step.

Theorem 2.1. *The adjoint method of a collocation method (Definition II.1.3) based on* $c_1,\ldots,c_s$ *is a collocation method based on* $c_1^*,\ldots,c_s^*$*, where*

$$c_i^* \;=\; 1 - c_{s+1-i}. \tag{2.1}$$

In the case that $c_i = 1 - c_{s+1-i}$ *for all i, the collocation method is symmetric.*

The adjoint method of a discontinuous collocation method (Definition II.1.7) based on b_1, b_s *and* $c_2,\ldots,c_{s-1}$ *is a discontinuous collocation method based on* b_1^*, b_s^* *and* $c_2^*,\ldots,c_{s-1}^*$*, where*

$$b_1^* = b_s, \quad b_s^* = b_1 \quad \text{and} \quad c_i^* = 1 - c_{s+1-i}. \tag{2.2}$$

In the case that $b_1 = b_s$ *and* $c_i = 1 - c_{s+1-i}$ *for all i, the discontinuous collocation method is symmetric.*

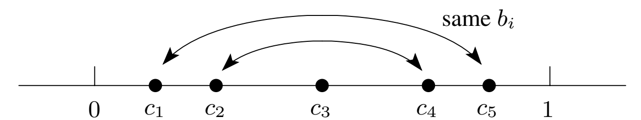


Fig. 2.1. Symmetry of collocation methods

Proof. Exchanging $(t_0, y_0) \leftrightarrow (t_1, y_1)$ and $h \leftrightarrow -h$ in the definition of a collocation method we get $u(t_1) = y_1$, $\dot u(t_1 - c_ih) = f\big(t_1 - c_ih, u(t_1 - c_ih)\big)$, and $y_0 = u(t_1 - h)$. Inserting $t_1 = t_0 + h$ this yields the collocation method based on c_i^* of (2.1). Observe that the c_i^* can be arbitrarily permuted. For discontinuous collocation methods the proof is similar. □

The preceding theorem immediately yields the following result.

Corollary 2.2. *The Gauss formulas (Table II.1.1), as well as the Lobatto IIIA (Table II.1.2) and Lobatto IIIB formulas (Table II.1.4) are symmetric integrators.* □

Theorem 2.3 (Stetter 1973, Wanner 1973). *The adjoint method of an s-stage Runge–Kutta method (II.1.4) is again an s-stage Runge–Kutta method. Its coefficients are given by*

$$a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}, \qquad b_i^* = b_{s+1-i}. \tag{2.3}$$

If

$$a_{s+1-i,s+1-j} + a_{ij} = b_j \qquad \textit{for all } i, j, \tag{2.4}$$

then the Runge–Kutta method (II.1.4) is symmetric.[2]

Proof. Exchanging $y_0 \leftrightarrow y_1$ and $h \leftrightarrow -h$ in the Runge–Kutta formulas yields

$$k_i = f\Big(y_0 + h\sum_{j=1}^{s}(b_j - a_{ij})k_j\Big), \qquad y_1 = y_0 + h\sum_{i=1}^{s} b_i k_i. \tag{2.5}$$

Since the values $\sum_{j=1}^{s}(b_j - a_{ij}) = 1 - c_i$ appear in reverse order, we replace k_i by k_{s+1-i} in (2.5), and then we substitute all indices i and j by $s+1-i$ and $s+1-j$, respectively. This proves (2.3).

The assumption (2.4) implies $a_{ij}^* = a_{ij}$ and $b_i^* = b_i$, so that $\Phi_h^* = \Phi_h$. □

Explicit Runge–Kutta methods cannot fulfill condition (2.4) with $i = j$, and it is not difficult to see that no explicit Runge–Kutta can be symmetric (Exercise 2). Let us therefore turn our attention to *diagonally implicit Runge–Kutta methods* (DIRK), for which $a_{ij} = 0$ for $i < j$, but with diagonal elements that can be non-zero. In this case condition (2.4) becomes

$$a_{ij} = b_j = b_{s+1-j} \quad \text{for } i > j, \qquad a_{jj} + a_{s+1-j,s+1-j} = b_j. \tag{2.6}$$

The Runge–Kutta tableau of such a method is thus of the form (e.g., for $s = 5$)

$$\begin{array}{c|ccccc}
c_1 & a_{11} \\
c_2 & b_1 & a_{22} \\
c_3 & b_1 & b_2 & a_{33} \\
1-c_2 & b_1 & b_2 & b_3 & a_{44} \\
1-c_1 & b_1 & b_2 & b_3 & b_2 & a_{55} \\
\hline
 & b_1 & b_2 & b_3 & b_2 & b_1
\end{array} \tag{2.7}$$

with $a_{33} = b_3/2$, $a_{44} = b_2 - a_{22}$, and $a_{55} = b_1 - a_{11}$. If one of the b_i vanishes, then the corresponding stage does not influence the numerical result. This stage can therefore be suppressed, so that the method is equivalent to one with fewer stages. Our next result shows that methods (2.7) can be interpreted as the composition of θ-methods, which are defined as

[2] For irreducible Runge–Kutta methods, the condition (2.4) is also necessary for symmetry (after a suitable permutation of the stages).

$$\Phi_h^\theta(y_0) = y_1, \qquad \text{where} \qquad y_1 = y_0 + hf\big((1-\theta)y_0 + \theta y_1\big). \tag{2.8}$$

Observe that the adjoint of the θ-method is $\Phi_h^{\theta *} = \Phi_h^{1-\theta}$.

Theorem 2.4. *A diagonally implicit Runge–Kutta method satisfying the symmetry condition (2.4) and* $b_i \neq 0$ *is equivalent to a composition of* θ*-methods*

$$\Phi_{b_1 h}^{\alpha_1 *} \circ \Phi_{b_2 h}^{\alpha_2 *} \circ \ldots \circ \Phi_{b_2 h}^{\alpha_2} \circ \Phi_{b_1 h}^{\alpha_1}, \tag{2.9}$$

where $\alpha_i = a_{ii}/b_i$.

Proof. Since the θ-method is a Runge–Kutta method with tableau

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array}$$

this follows from the discussion in Sect. III.1.3. We have used $\Phi_{b_{s+1-i}h}^{\alpha_{s+1-i}} = \Phi_{b_i h}^{\alpha_i *}$ which holds, because $b_{s+1-i} = b_i$ and $\alpha_{s+1-i} = 1 - \alpha_i$ by (2.6). □

A more detailed discussion of such methods is therefore postponed to Sect. V.3 on symmetric composition methods.

V.2.2 Partitioned Runge–Kutta Methods

Applying partitioned Runge–Kutta methods (II.2.2) to general partitioned systems

$$\dot y = f(y,z), \qquad \dot z = g(y,z), \tag{2.10}$$

it is obvious that for their symmetry both Runge–Kutta methods have to be symmetric (because $\dot y = f(y)$ and $\dot z = g(z)$ are special cases of (2.10)). The proof of the following result is identical to that of Theorem 2.3 and therefore omitted.

Theorem 2.5. *If the coefficients of both Runge–Kutta methods* b_i, a_{ij} *and* $\widehat b_i, \widehat a_{ij}$ *satisfy the condition (2.4), then the partitioned Runge–Kutta method (II.2.2) is symmetric.* □

As a consequence of this theorem we obtain that the Lobatto IIIA-IIIB pair (see Sect. II.2.2) and, in particular, the Störmer–Verlet scheme are symmetric integrators.

An interesting feature of partitioned Runge–Kutta methods is the possibility of having *explicit, symmetric* methods for problems of the form

$$\dot y = f(z), \qquad \dot z = g(y). \tag{2.11}$$

Second order differential equations $\ddot y = g(y)$, written in the form $\dot y = z, \dot z = g(y)$ have this structure, and also all Hamiltonian systems with separable Hamiltonian $H(p,q) = T(p) + V(q)$. It is not possible to get explicit symmetric integrators with non-partitioned Runge–Kutta methods (Exercise 2).

The Störmer–Verlet method (Table II.2.1) applied to (2.11) reads

$$\begin{aligned} z_{1/2} &= z_0 + \frac{h}{2}\, g(y_0) \\ y_1 &= y_0 + h\, f(z_{1/2}) \\ z_1 &= z_{1/2} + \frac{h}{2}\, g(y_1) \end{aligned}$$

and is the composition $\Phi^*_{h/2} \circ \Phi_{h/2}$, where

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \Phi_h \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}, \qquad \begin{aligned} y_1 &= y_0 + hf(z_1) \\ z_1 &= z_0 + hg(y_0) \end{aligned} \tag{2.12}$$

is the symplectic Euler method and

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \Phi^*_h \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}, \qquad \begin{aligned} y_1 &= y_0 + hf(z_0) \\ z_1 &= z_0 + hg(y_1) \end{aligned} \tag{2.13}$$

its adjoint. All these methods are obviously explicit. How can they be extended to higher order? The idea is to consider partitioned Runge–Kutta methods based on diagonally implicit methods such as in (2.7). If $a_{ii} \cdot \widehat{a}_{ii} = 0$, then one component of the ith stage is given explicitly and, due to the special structure of (2.11), the other component is also obtained in a straightforward manner. In order to achieve $a_{ii} \cdot \widehat{a}_{ii} = 0$ with a symmetric partitioned method, we have to assume that s, the number of stages, is even.

Theorem 2.6. *A partitioned Runge–Kutta method, based on two diagonally implicit methods satisfying $a_{ii} \cdot \widehat{a}_{ii} = 0$ and (2.4) with $b_i \neq 0$ and $\widehat{b}_i \neq 0$, is equivalent to a composition of $\Phi_{b_i h}$ and $\Phi^*_{b_i h}$ with Φ_h and Φ^*_h given by (2.12) and (2.13), respectively.* □

For example, the partitioned method

$$\begin{array}{c|cccc} & 0 & & & \\ & b_1 & b_2 & & \\ & b_1 & b_2 & 0 & \\ & b_1 & b_2 & b_2 & b_1 \\ \hline & b_1 & b_2 & b_2 & b_1 \end{array} \qquad \begin{array}{c|cccc} & \widehat{b}_1 & & & \\ & \widehat{b}_1 & 0 & & \\ & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_2 & \\ & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_2 & 0 \\ \hline & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_2 & \widehat{b}_1 \end{array}$$

satisfies the assumptions of the preceding theorem. Since the methods have identical stages, the numerical result only depends on $\widehat{b}_1$, $b_1 + b_2$, $\widehat{b}_2 + \widehat{b}_3$, $b_3 + b_4$, and $\widehat{b}_4$. Therefore, we can assume that $\widehat{b}_i = b_i$ and the method is equivalent to the composition $\Phi^*_{b_1 h} \circ \Phi_{b_2 h} \circ \Phi^*_{b_2 h} \circ \Phi_{b_1 h}$.

V.3 Symmetric Composition Methods

In Sect. II.4 the idea of composition methods is introduced, and a systematic way of obtaining high-order methods is outlined. These methods, based on (II.4.4) or on

(II.4.5), turn out to be symmetric, but they require too many stages. A theory of order conditions for general composition methods is developed in Sect. III.3. Here, we apply this theory to the construction of high-order symmetric methods. We mainly follow two lines.

- *Symmetric composition of first order methods.*

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi^*_{\beta_s h} \circ \ldots \circ \Phi^*_{\beta_2 h} \circ \Phi_{\alpha_1 h} \circ \Phi^*_{\beta_1 h}, \tag{3.1}$$

  where Φ_h is an arbitrary first order method. In order to make this method symmetric, we assume $\alpha_s = \beta_1$, $\alpha_{s-1} = \beta_2$, etc.
- *Symmetric composition of symmetric methods.*

$$\Psi_h = \Phi_{\gamma_s h} \circ \Phi_{\gamma_{s-1} h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}, \tag{3.2}$$

  where Φ_h is a symmetric second order method and $\gamma_s = \gamma_1$, $\gamma_{s-1} = \gamma_2$, etc.

V.3.1 Symmetric Composition of First Order Methods

Because of Lemma 3.2 below, every method (3.2) is a special case of method (3.1). In this subsection we concentrate on methods that are of the form (3.1) but not of the form (3.2).

For constructing methods (3.1) of a certain order, one has to solve the system of nonlinear equations given in Theorem III.3.14 (see also Example III.3.15). The symmetry assumption on the coefficients considerably simplifies this system.

Theorem 3.1. *If the coefficients of method (3.1) satisfy $\alpha_{s+1-i} = \beta_i$ for all i, then it is sufficient to consider those trees with odd $\|\tau\|$.*

Proof. This is a consequence of Theorem II.3.2 (the maximal order of symmetric methods is even). In fact, if the condition for order 1 is satisfied, it is automatically of order 2. If, in addition, the conditions for order 3 are satisfied, it is automatically of order 4, etc. □

> It may come as a surprise that the popular leapfrog . . . can be beaten, but only slightly. (R.I. McLachlan 1995)

Methods of Order 2. The only remaining condition for order two is $\sum_{k=1}^{s}(\alpha_k + \beta_k) = 1$, and, for $s = 1$, the symmetry requirement leads to $\Phi_{h/2} \circ \Phi^*_{h/2}$. Depending on the choice of Φ_h, this method is equivalent to the midpoint rule, the trapezoidal rule, or the Störmer–Verlet scheme, all very famous and frequently used. However, McLachlan (1995) discovered that the case $s = 2$ can be slightly more advantageous. We obtain

$$\Phi_{\alpha h} \circ \Phi^*_{(1/2-\alpha)h} \circ \Phi_{(1/2-\alpha)h} \circ \Phi^*_{\alpha h}, \tag{3.3}$$

where α is a free parameter, which can serve for clever tuning.

Minimizing the Local Error of Composition Methods. Subtracting the B_∞-series of the numerical and the exact solutions (see Sect. III.3.2), we obtain

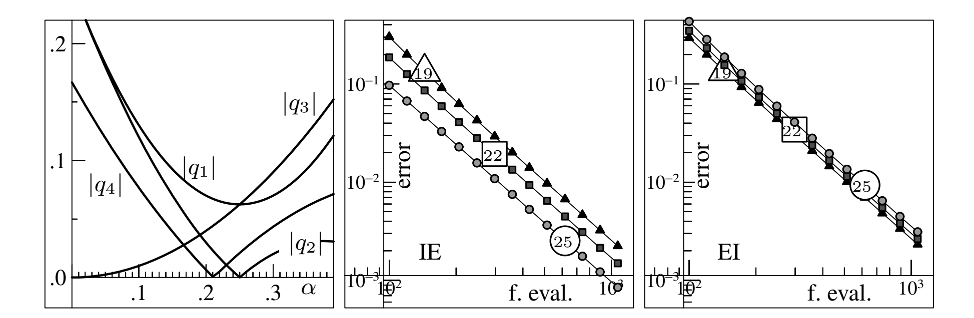


Fig. 3.1. The error functions $|q_i(\alpha)|$ defined in (3.5) (left picture). Work-precision diagrams for the Kepler problem (as in Fig. II.4.4) and for method (3.3) with $\alpha = 0.25$ (Störmer–Verlet), $\alpha = 0.1932$ (McLachlan), and $\alpha = 0.22$. "IE": method Φ_h treats position by implicit Euler, velocity by explicit Euler; "EI": method Φ_h treats position by explicit Euler, velocity by implicit Euler

$$\Psi_h(y) - \varphi_h(y) \;=\; h^{p+1} \sum_{\|\tau\|=p+1} \frac{1}{\sigma(\tau)} \Big(a(\tau) - \mathrm{e}(\tau) \Big)\, F(\tau)(y) + \mathcal{O}(h^{p+2}).$$

Assuming that the basic method has an expansion $\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \ldots$, we obtain for method (3.3), similar to (III.3.3), the local error

$$h^3 \Big(q_1(\alpha) d_3(y) + q_2(\alpha)\big(d_2' f\big)(y) + q_3(\alpha)\big(f' d_2\big)(y) + \frac{1}{2}\, q_4(\alpha)\big(f''(f,f)\big)(y) + q_5(\alpha)\big(f' f' f\big)(y)\Big) + \mathcal{O}(h^4), \tag{3.4}$$

which contains one term for each of the 5 trees $\tau \in T_\infty$ with $||\tau|| = 3$. The $q_i(\alpha)$ are the polynomials

$$q_1(\alpha) = \frac{1}{4}\big(1 - 6\alpha + 12\alpha^2\big), \qquad q_2(\alpha) = \frac{1}{4}\big(-1 + 6\alpha - 8\alpha^2\big),$$
$$q_3(\alpha) = -\alpha^2, \qquad q_4(\alpha) = \frac{1}{6}\big(1 - 6\alpha + 6\alpha^2\big), \qquad q_5(\alpha) = \frac{1}{3} q_1(\alpha), \tag{3.5}$$

which are plotted in the left picture of Fig. 3.1. If we allow arbitrary basic methods and arbitrary problems, all elementary differentials in the local error are independent, and there is no overall optimal value for α. We see that the modulus of $q_1(\alpha)$ and $q_2(\alpha)$ are minimal for $\alpha = 1/4$, which is precisely the value corresponding to a double application of $\Phi_{h/2} \circ \Phi^*_{h/2}$ with halved step size. But the values $|q_3(\alpha)|$ and $|q_4(\alpha)|$ become smaller with decreasing α (close to $\alpha = 1/4$). McLachlan (1995) therefore minimizes some norm of the error (see Exercise 4) and arrives at the value $\alpha = 0.1932$.

In the numerical experiment of Fig. 3.1 we apply method (3.3) with three different values of α to the Kepler problem (with data as in Fig. II.4.4 and the symplectic Euler method for Φ_h). Once we treat the position variable by the implicit Euler method and the velocity variable by the explicit Euler method (central picture), and

once the other way round (right picture). We notice that the method which is best in one case is worst in the other.

This simple experiment shows that choosing the free parameters of the method by minimizing some arbitrary measure of the error coefficients is problematic. For higher order methods there are many more expressions in the dominating term of the local error (for example: 29 terms for $||\tau|| = 5$). The corresponding functions q_i give a lot of information on the local error, and they indicate the region of parameters that produce good methods. But, unless more information is known about the problem (second order differential equation, nearly integrable systems), one usually minimizes, for orders of 8 or 10, just the maximal values of the α_i, β_i, or γ_i (Kahan & Li 1997).

Methods of Order 4. Theorem 3.1 and Example III.3.15 give 3 conditions for order 4. Therefore, we put $s = 3$ in (3.1) and assume symmetry $\beta_1 = \alpha_3$, $\beta_2 = \alpha_2$, and $\beta_3 = \alpha_1$. This leads to the conditions

$$\alpha_1 + \alpha_2 + \alpha_3 = \frac{1}{2}, \qquad \alpha_1^3 + \alpha_2^3 + \alpha_3^3 = 0, \qquad (\alpha_3^2 - \alpha_1^2)(\alpha_1 + \alpha_2) = 0.$$

Since with $\alpha_1 + \alpha_2 = 0$ or with $\alpha_1 + \alpha_3 = 0$ the first two of these equations are not compatible, the unique solution of this system is

$$\alpha_1 = \alpha_3 = \frac{1}{2\,(2 - 2^{1/3})}, \quad \alpha_2 = -\frac{2^{1/3}}{2\,(2 - 2^{1/3})}.$$

We observe that $\beta_i = \alpha_i$ for all i. Therefore, $\Phi_{\alpha_i h} \circ \Phi^*_{\beta_i h}$ can be grouped together in (3.1) and we have obtained a method of type (3.2), which is actually method (II.4.4) with $p = 2$.

Again, the solutions with the minimal number of stages do not give the best methods (remember the good performance of Suzuki's fourth order method (II.4.5) in Fig. II.4.4), so we look for 4th order methods with larger s. McLachlan (1995) has constructed a method for $s = 5$ with particularly small error terms and nice coefficients

$$\begin{aligned} &\beta_1 = \alpha_5 = \frac{14 - \sqrt{19}}{108}, \qquad \alpha_1 = \beta_5 = \frac{146 + 5\sqrt{19}}{540}, \\ &\beta_2 = \alpha_4 = \frac{-23 - 20\sqrt{19}}{270}, \quad \alpha_2 = \beta_4 = \frac{-2 + 10\sqrt{19}}{135}, \quad \beta_3 = \alpha_3 = \frac{1}{5}, \end{aligned} \tag{3.6}$$

which he recommends "for all uses".

In Fig. 3.2 we compare the numerical performances of all these methods on our already well-known example in both variants (implicit-explicit and vice-versa). We see that the best methods in *one* picture may be worse in the other. For comparison, the results are surrounded by "ghosts in grey" representing good formulae from the next lower (order 2) and the next higher (order 6) class of methods.

Methods Tuned for Special Problems. In the case where one is applying a *special* method to a *special* problem (e.g., to second order differential equations or to small

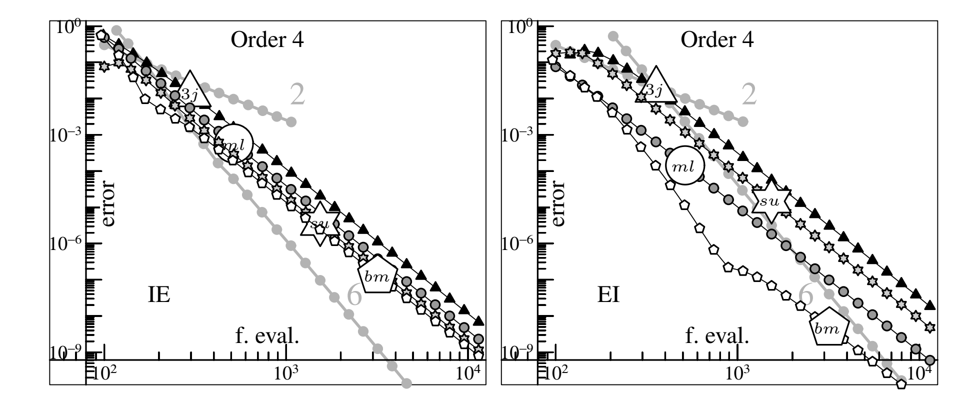


Fig. 3.2. Work-precision diagrams for methods of order 4 as in Fig. 3.1; "3j": the Triple Jump (II.4.4); "su": method (II.4.5) of Suzuki; "ml": McLachlan (3.6); "bm": method (3.7); in grey: neighbouring order methods Störmer/Verlet (order 2) and $p6\,s9$ (order 6)

perturbations of integrable systems), more spectacular gains of efficiency are possible. For example, Blanes & Moan (2002) have constructed the following fourth order method with $s = 6$

$$\begin{aligned}
\beta_1 = \alpha_6 &= 0.082984406417405, & \alpha_1 = \beta_6 &= 0.16231455076687,\\
\beta_2 = \alpha_5 &= 0.23399525073150, & \alpha_2 = \beta_5 &= 0.37087741497958\,,\\
\beta_3 = \alpha_4 &= -0.40993371990193, & \alpha_3 = \beta_4 &= 0.059762097006575\,,
\end{aligned} \tag{3.7}$$

which, when correctly applied to second order differential equations (right picture of Fig. 3.2) exhibits excellent performance.

Further methods, adapted to the integration of second order differential equations, have been constructed by Forest (1992), McLachlan & Atela (1992), Calvo & Sanz-Serna (1993), Okunbor & Skeel (1994), and McLachlan (1995). Another important situation, which allows a tuning of the parameters, are near-integrable systems such as the perturbed two-body motion (e.g., the outer solar system considered in Chap. I). If the differential equation can be split into $\dot y = f^{[1]}(y) + f^{[2]}(y)$, where $\dot y = f^{[1]}(y)$ is exactly integrable and $f^{[2]}(y)$ is small compared to $f^{[1]}(y)$, special integrators should be used. We refer to Kinoshita, Yoshida & Nakai (1991), Wisdom & Holman (1991), Saha & Tremaine (1992), and McLachlan (1995b) for more details and for the parameters of such integrators.

Methods of Order 6. By Theorem 3.1 and Example III.3.12 a method (3.1) has to satisfy 9 conditions for order 6. It turns out that these order conditions have already a solution with $s = 7$, but all known solutions with $s \le 8$ are equivalent to methods of type (3.2). With order 6 we are apparently close to the point where the enormous simplifications of the order conditions due to Theorem 3.3 below start to outperform the freedom of choosing different values for α_i and β_i. We therefore continue our discussion by considering only the special case (3.2).

V.3.2 Symmetric Composition of Symmetric Methods

The introduction of more symmetries into the method simplifies considerably the order conditions. These simplifications can be best understood with a sort of "Choleski decomposition" of symmetric methods (Murua & Sanz-Serna 1999).

Lemma 3.2. *For every symmetric method $\Phi_h(y)$ that admits an expansion in powers of h, there exists $\widehat{\Phi}_h(y)$ such that*

$$\Phi_h(y) = \big(\widehat{\Phi}_{h/2} \circ \widehat{\Phi}^*_{h/2}\big)(y).$$

Proof. Since $\Phi_h(y) = y + \mathcal{O}(h)$ is close to the identity, the existence of a unique method $\widehat{\Phi}_h(y) = y + hd_1(y) + h^2 d_2(y) + \ldots$ satisfying $\Phi_h = \widehat{\Phi}_{h/2} \circ \widehat{\Phi}_{h/2}$ follows from Taylor expansion and from a comparison of like powers of h.

If $\Phi_h(y)$ is symmetric, we have in addition

$$\Phi_h = \Phi^{-1}_{-h} = \widehat{\Phi}^{-1}_{-h/2} \circ \widehat{\Phi}^{-1}_{-h/2},$$

and $\widehat{\Phi}_{h/2} = \widehat{\Phi}^{-1}_{-h/2} = \widehat{\Phi}^*_{h/2}$ follows from the uniqueness of $\widehat{\Phi}_h$. □

We let Φ_h be a symmetric method, and we consider the composition

$$\Psi_h = \Phi_{\gamma_s h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}. \tag{3.8}$$

Using the method $\widehat{\Phi}_h$ of Lemma 3.2, this composition method is equivalent to (3.1) (Φ_h replaced with $\widehat{\Phi}_h$) with

$$\alpha_i = \beta_i = \frac{\gamma_i}{2}. \tag{3.9}$$

Theorem 3.3. *For composition methods (3.8) with symmetric Φ_h it is sufficient to consider the order conditions of Theorem III.3.14 for $\tau \in \mathcal{H}$ where all vertices of τ have odd indices.*

Proof. If $i(\tau)$ is even, it follows from $\alpha_k = \beta_k$ and from (III.3.11) that

$$a_s(\tau) = a_{s-1}(\tau) = \ldots = a_1(\tau) = a_0(\tau) = 0.$$

Since $\mathbf{e}(\tau) = 0$ for such trees, the corresponding order condition is automatically satisfied. Any other vertex with an even index can be brought to the root by applying the Switching Lemma III.3.8. □

After this reduction, only 7 conditions survive for order 6 from the trees displayed in Example III.3.12. A further reduction in the number of order conditions is achieved by assuming *symmetric coefficients* in method (3.8), i.e.,

$$\gamma_{s+1-j} = \gamma_j \qquad \text{for all } j. \tag{3.10}$$

This implies that the overall method Ψ_h is symmetric, so that the order conditions for trees with an even $||\tau||$ need not be considered. This proves the following result.

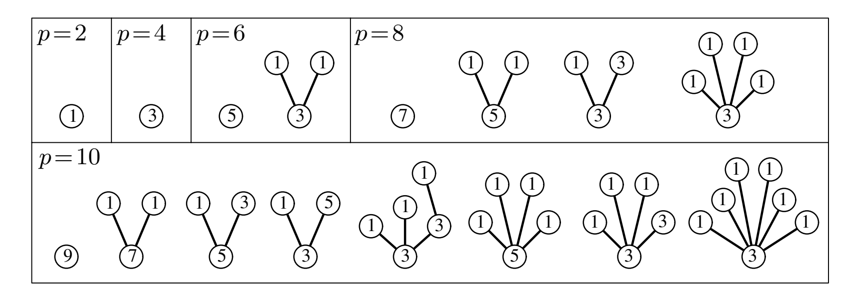


Fig. 3.3. Symmetric Composition of Symmetric Methods up to order 10

Theorem 3.4. *For composition methods (3.8) with symmetric Φ_h, satisfying (3.10), it is sufficient to consider the order conditions for $\tau \in \mathcal{H}$ where all vertices of τ have odd indices and where $\|\tau\|$ is odd.* □

Figure 3.3 shows the remaining order conditions for methods up to order 10. We see that for order 6 there remain only 4 conditions, much less than the 166 that we started with (Theorem III.3.6).

Example 3.5. The rule of (III.3.14) leads to the following conditions for *symmetric* composition of *symmetric* methods:

Order 2:	①	$\sum_{k=1}^{s} \gamma_k = 1$		
Order 4:	③	$\sum_{k=1}^{s} \gamma_k^3 = 0$		
Order 6:	⑤	$\sum_{k=1}^{s} \gamma_k^5 = 0$		$\sum_{k=1}^{s} \gamma_k^3 \Big({\sum_{\ell=1}^{k}}' \gamma_\ell \Big)^2 = 0$
Order 8:	⑦	$\sum_{k=1}^{s} \gamma_k^7 = 0$		$\sum_{k=1}^{s} \gamma_k^5 \Big({\sum_{\ell=1}^{k}}' \gamma_\ell \Big)^2 = 0$
		$\sum_{k=1}^{s} \gamma_k^3 {\sum_{\ell=1}^{k}}' \gamma_\ell {\sum_{m=1}^{k}}' \gamma_m^3 = 0$		$\sum_{k=1}^{s} \gamma_k^3 \Big({\sum_{\ell=1}^{k}}' \gamma_\ell \Big)^4 = 0.$

Here, similar to Example III.3.15, a *prime* attached to a summation symbol indicates that the last term γ_ℓ^i is taken as $\gamma_\ell^i/2$.

Methods of Order 4. The methods (II.4.4) and (II.4.5) are both of the form (3.8), and those with $p = 2$ yield methods of order 4. We have seen in the experiment of Fig. II.4.4 that the method (II.4.5) yields more precise approximations; see also Fig. 3.2. We do not know of any 4th order method of type (3.2) that is significantly better than method (3.1) with coefficients (3.6).

Methods of Order 6. If we search for a minimal stage solution of the four equations for order 6, we apparently need four free parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$; then $\gamma_5, \gamma_6, \gamma_7$ are determined by symmetry. Equation ① gives $\gamma_4 = 1-2(\gamma_1+\gamma_2+\gamma_3)$. So we end up with three equations for the three unknowns $\gamma_1, \gamma_2, \gamma_3$. A numerical search for this problem produces three solutions, the best of which has been discovered by many authors, in particular by Yoshida (1990), and is as follows:

$$\begin{aligned} \gamma_1 = \gamma_7 &= 0.78451361047755726381949763 \\ \gamma_2 = \gamma_6 &= 0.23557321335935813368479318 \\ \gamma_3 = \gamma_5 &= -1.17767998417887100694641568 \\ \gamma_4 &= 1.31518632068391121888424973 \end{aligned} \qquad p6\,s7 \tag{3.11}$$

Using computer algebra, Koseleff (1996) proves that the nonlinear system for $\gamma_1, \gamma_2, \gamma_3$ has not more than three real solutions.

Similar to the situation for order 4, where relaxing the minimal number of stages allowed a significant increase of performance, we also might expect to obtain better methods of order 6 in this way. McLachlan (1995) increases s by two and constructs good methods with small error coefficients. By minimizing $\max_i |\gamma_i|$, Kahan & Li (1997) obtain the following excellent method [3]

$$\begin{aligned} \gamma_1 = \gamma_9 &= 0.39216144400731413927925056 \\ \gamma_2 = \gamma_8 &= 0.33259913678935943859974864 \\ \gamma_3 = \gamma_7 &= -0.70624617255763935980996482 \\ \gamma_4 = \gamma_6 &= 0.08221359629355080023149045 \\ \gamma_5 &= 0.79854399093482996339895035 \end{aligned} \qquad p6\,s9 \tag{3.12}$$

This method produces, with a comparable number of total steps, errors which are typically smaller than those of method (3.11). Numerical results of these two methods are given in Fig. 3.4.

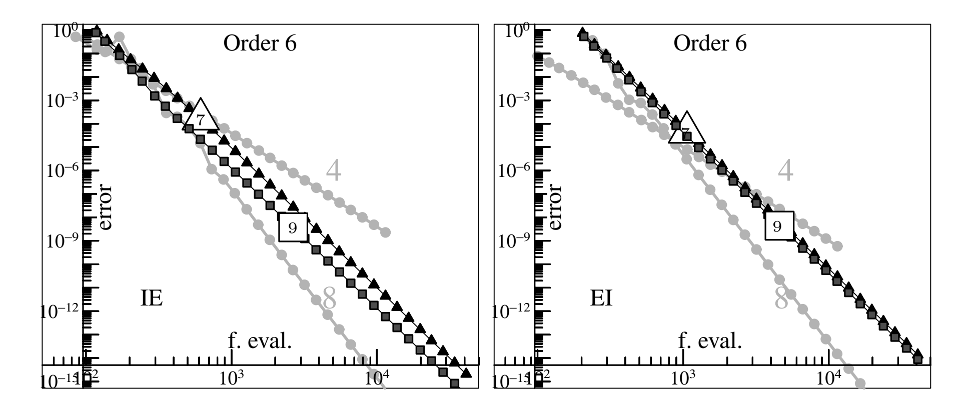


Fig. 3.4. Work-precision diagrams for methods of order 6 for the Kepler problem as in Fig. 3.1; "7": method $p6\,s7$ of (3.11); "9": method $p6\,s9$ of (3.12); in grey: neighbouring order methods (3.6) (order 4) and $p8\,s17$ (order 8)

[3] The authors are grateful to S. Blanes for this reference.

Methods of Order 8. For order 8, Fig. 3.3 represents 8 equations to solve. This indicates that the minimal value of s is 15. A numerical search for solutions $\gamma_1, \ldots, \gamma_8$ of these equations produces hundreds of solutions. We choose among all these the solution with the smallest $\max(|\gamma_i|)$. The coefficients, which were originally given by Suzuki & Umeno (1993), Suzuki (1994), and later by McLachlan (1995), are as follows:

$$
\begin{aligned}
\gamma_1 = \gamma_{15} &= 0.74167036435061295344822780 \\
\gamma_2 = \gamma_{14} &= -0.40910082580003159399730010 \\
\gamma_3 = \gamma_{13} &= 0.19075471029623837995387626 \\
\gamma_4 = \gamma_{12} &= -0.57386247111608226665638773 \\
\gamma_5 = \gamma_{11} &= 0.29906418130365592384446354 \\
\gamma_6 = \gamma_{10} &= 0.33462491824529818378495798 \\
\gamma_7 = \gamma_9 &= 0.31529309239676659663205666 \\
\gamma_8 &= -0.79688793935291635401978884
\end{aligned}
\qquad p8\,s15 \tag{3.13}
$$

By putting $s = 17$ we obtain one degree of freedom in solving the equations. This allows an improvement on the foregoing method. The best known solution, slightly better than a method of McLachlan (1995), has been found by Kahan & Li (1997) and is given by

$$
\begin{aligned}
\gamma_1 = \gamma_{17} &= 0.13020248308889008087881763 \\
\gamma_2 = \gamma_{16} &= 0.56116298177510838456196441 \\
\gamma_3 = \gamma_{15} &= -0.38947496264484728640807860 \\
\gamma_4 = \gamma_{14} &= 0.15884190655515560089621075 \\
\gamma_5 = \gamma_{13} &= -0.39590389413323757733623154 \\
\gamma_6 = \gamma_{12} &= 0.18453964097831570709183254 \\
\gamma_7 = \gamma_{11} &= 0.25837438768632204729397911 \\
\gamma_8 = \gamma_{10} &= 0.29501172360931029887096624 \\
\gamma_9 &= -0.60550853383003451169892108
\end{aligned}
\qquad p8\,s17 \tag{3.14}
$$

Numerical results, in the same style as above, are given in Fig. 3.5.

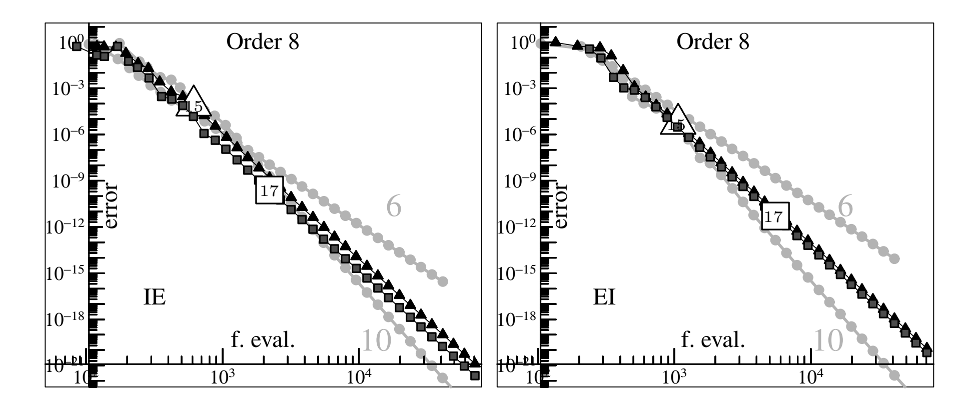


Fig. 3.5. Work-precision diagrams for methods of order 8 for the Kepler problem as in Fig. 3.1; "15": method $p8\,s15$ of (3.13); "17": method $p8\,s17$ of (3.14); in grey: neighbouring order methods $p6\,s9$ (order 6) and $p10\,s35$ (order 10)

Methods of Order 10. The first methods of order 10 were given by Kahan & Li (1997) with $s = 31$ and $s = 33$, which could be improved on after some nights of computer search (see method (V.3.15) of the first edition). A significantly improved method for $s = 35$ (see Fig. 3.5 for a comparison with eighth order methods) has in the meantime been found by Sofroniou & Spaletta (2004):

$$
\begin{aligned}
\gamma_1 = \gamma_{35} &= 0.078795722521686419263907679 \\
\end{aligned}
$$

$$
\begin{aligned}
\gamma_1 = \gamma_{35} &= 0.07879572252168641926390768 \\
\gamma_2 = \gamma_{34} &= 0.31309610341510852776481247 \\
\gamma_3 = \gamma_{33} &= 0.02791838323507806610952027 \\
\gamma_4 = \gamma_{32} &= -0.22959284159390709415121340 \\
\gamma_5 = \gamma_{31} &= 0.13096206107716486317465686 \\
\gamma_6 = \gamma_{30} &= -0.26973340565451071434460973 \\
\gamma_7 = \gamma_{29} &= 0.07497334315589143566613711 \\
\gamma_8 = \gamma_{28} &= 0.11199342399981020488957508 \\
\gamma_9 = \gamma_{27} &= 0.36613344954622675119314812 \\
\gamma_{10} = \gamma_{26} &= -0.39910563013603589787862981 \\
\gamma_{11} = \gamma_{25} &= 0.10308739852747107731580277 \\
\gamma_{12} = \gamma_{24} &= 0.41143087395589023782070412 \\
\gamma_{13} = \gamma_{23} &= -0.00486636058313526176219566 \\
\gamma_{14} = \gamma_{22} &= -0.39203335370863990644808194 \\
\gamma_{15} = \gamma_{21} &= 0.05194250296244964703718290 \\
\gamma_{16} = \gamma_{20} &= 0.05066509075992449633587434 \\
\gamma_{17} = \gamma_{19} &= 0.04967437063972987905456880 \\
\gamma_{18} &= 0.04931773575959453791768001
\end{aligned}
\qquad (3.15)
$$

$p10\, s35$

0 1

V.3.3 Effective Order and Processing Methods

> There has recently been a revival of interest in the concept of "effective order". (J.C. Butcher 1998)

The concept of effective order was introduced by Butcher (1969) with the aim of constructing 5th order explicit Runge–Kutta methods with 5 stages. The idea is to search for a computationally efficient method K_h such that with a suitable χ_h,

$$\Psi_h \;=\; \chi_h \circ K_h \circ \chi_h^{-1} \qquad (3.16)$$

has an order higher than that of K_h. The method K_h is called the *kernel*, and χ_h can be interpreted as a transformation in the phase space, close to the identity. Because of

$$\Psi_h^N \;=\; \chi_h \circ K_h^N \circ \chi_h^{-1},$$

an implementation of Ψ_h over N steps with constant step size h has the same computational efficiency as K_h. The computation of χ_h^{-1} has only to be done once at the beginning of the integration, and χ_h has to be evaluated only at output points, which can be performed on another processor. In the article López-Marcos, Sanz-Serna & Skeel (1996) the notion of *preprocessing* for the step χ_h^{-1} and *postprocessing* for χ_h is introduced.

Example 3.6 (Störmer–Verlet as Processed Symplectic Euler Method). Consider a split differential equation, let $\Phi_h^{[LT]} = \varphi_h^{[2]} \circ \varphi_h^{[1]}$ be the Lie-Trotter formula or symplectic Euler method (see Sect. II.5), and $\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ the Strang splitting or Störmer–Verlet scheme. As a consequence of the group property of the exact flow, we have

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \Phi_h^{[LT]} \circ \varphi_{-h/2}^{[1]} = \chi_h \circ \Phi_h^{[LT]} \circ \chi_h^{-1}$$

with $\chi_h = \varphi_{h/2}^{[1]}$. Hence, applying the Lie-Trotter formula with processing yields a second order approximation.

Since the use of geometric integrators requires constant step sizes, it is quite natural that Butcher's idea of effective order has been revived in this context. A systematic search for processed composition methods started with the works of Wisdom, Holman & Touma (1996), McLachlan (1996), and Blanes, Casas & Ros (1999, 2000b).

Let us explain the technique of processing in the situation where the kernel K_h is a symmetric composition

$$K_h = \Phi_{\gamma_s h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h} \qquad (\gamma_{s+1-i} = \gamma_i \text{ for all } i) \tag{3.17}$$

of a symmetric method Φ_h. We suppose that the processor is of the form

$$\chi_h = \Phi_{\delta_r h} \circ \ldots \circ \Phi_{\delta_2 h} \circ \Phi_{\delta_1 h}, \tag{3.18}$$

such that its inverse is given by (use the symmetry $\Phi_h^{-1} = \Phi_{-h}$)

$$\chi_h^{-1} = \Phi_{-\delta_1 h} \circ \Phi_{-\delta_2 h} \circ \ldots \circ \Phi_{-\delta_r h}. \tag{3.19}$$

Order Conditions. The composite method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is of the form $\Psi_h = \Phi_{\varepsilon_{2r+s} h} \circ \ldots \circ \Phi_{\varepsilon_2 h} \circ \Phi_{\varepsilon_1 h}$ with

$$(\varepsilon_{2r+s}, \ldots, \varepsilon_2, \varepsilon_1) = (\delta_r, \ldots, \delta_1, \gamma_s, \ldots, \gamma_1, -\delta_1, \ldots, -\delta_r). \tag{3.20}$$

Theorem 3.3 thus tells us that only the order conditions corresponding to $\tau \in \mathcal{H}$, whose vertices have odd indices, have to be considered. Unfortunately, the sequence $\{\varepsilon_i\}$ of (3.20) does not satisfy the symmetry relation (3.10), unless all δ_i vanish. However, if we require

$$\chi_{-h}(y) = \chi_h(y) + \mathcal{O}(h^{p+1}), \tag{3.21}$$

we see that $\chi_h^{-1}(y) = \chi_h^*(y) + \mathcal{O}(h^{p+1})$, and the method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is symmetric up to terms of order $\mathcal{O}(h^{p+1})$. Consequently, the reduction of Theorem 3.4 is valid, so that for order p only the trees of Fig. 3.3 have to be considered.

For the first tree of Example 3.5 the order condition is

$$1 = \sum_{k=1}^{2r+s} \varepsilon_k = \sum_{k=1}^{s} \gamma_k,$$

and we see that this is a condition on the kernel K_h only. Similarly, for odd i we have

$$0 = \sum_{k=1}^{2r+s} \varepsilon_k^i = \sum_{k=1}^{s} \gamma_k^i, \tag{3.22}$$

so that also the trees ③, ⑤, ⑦, ... give conditions on K_h and cannot be influenced by the processor. We next consider the trees of Example 3.5 with three vertices, whose order condition is

$$0 = \sum_{k=1}^{2r+s} \varepsilon_k^i \sum_{\ell=1}^{k}{}' \varepsilon_\ell^j \sum_{m=1}^{k}{}' \varepsilon_m^q .$$

We split the sums according to the partitioning into $\delta_i, \gamma_i, -\delta_i$ in (3.20), and we denote the expressions appearing in Example 3.5 by $a(\tau)$ and those corresponding to χ_h and χ_h^{-1} by $b(\tau)$ and $b^{-1}(\tau)$, respectively. Using the abbreviations τ_i for the tree with one vertex labelled i, τ_{ij} for the tree with two vertices labelled i (the root) and j, and by τ_{ijq} the trees with three vertices labelled i (root), j and q (vertices that are directly connected to the root), this yields

$$\begin{aligned} 0 \;=\;& b^{-1}(\tau_{ijq}) + a(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) + a(\tau_{ij})b^{-1}(\tau_q) \\ +\;& a(\tau_{iq})b^{-1}(\tau_j) + a(\tau_{ijq}) + b(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) \\ +\;& b(\tau_i)b^{-1}(\tau_j)a(\tau_q) + b(\tau_i)a(\tau_j)b^{-1}(\tau_q) + b(\tau_i)a(\tau_j)a(\tau_q) \\ +\;& b(\tau_{ij})b^{-1}(\tau_q) + b(\tau_{ij})a(\tau_q) + b(\tau_{iq})b^{-1}(\tau_j) + b(\tau_{iq})a(\tau_j) + b(\tau_{ijq}). \end{aligned} \tag{3.23}$$

How can we simplify this long expression? First of all, we imagine K_h to be the identity (either $s = 0$ or all $\gamma_i = 0$), so that $\Psi_h = \chi_h \circ \chi_h^{-1}$ becomes the identity. In this situation, the terms involving $a(\tau)$ are not present in (3.23), and we obtain

$$0 = b^{-1}(\tau_{ijq}) + b(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) + b(\tau_{ij})b^{-1}(\tau_q) + b(\tau_{iq})b^{-1}(\tau_j) + b(\tau_{ijq}).$$

We can thus remove all terms in (3.23) that do not contain a factor $a(\tau)$. Now observe that by (3.21), $\chi_h(y)$ as well as $\chi_h^{-1}(y)$ have an expansion in even powers of h. Therefore, $b(\tau)$ and $b^{-1}(\tau)$ vanish for all τ with odd $\|\tau\|$. Formula (3.23) thus simplifies considerably and yields

$$0 \;=\; a(\tau_{311}) + 2b(\tau_{31})a(\tau_1), \tag{3.24}$$

$$0 \;=\; a(\tau_{511}) + 2b(\tau_{51})a(\tau_1), \tag{3.25}$$

$$0 \;=\; a(\tau_{313}) + b(\tau_{31})a(\tau_3) + b(\tau_{33})a(\tau_1). \tag{3.26}$$

A similar computation for the last tree in Example 3.5 gives (in an obvious notation)

$$0 \;=\; a(\tau_{31111}) + 4b(\tau_{31})a(\tau_1)^3 + 4b(\tau_{3111})a(\tau_1). \tag{3.27}$$

Since $a(\tau_1) = \sum_{i=1}^{s} \gamma_i = 1$, the conditions (3.24), (3.25) and (3.27) can be interpreted as conditions on the processor, namely on $b(\tau_{31})$, $b(\tau_{51})$ and $b(\tau_{3111})$. We

already have $a(\tau_3) = 0$ from (3.22), and an application of the Switching Lemma III.3.8 gives $b(\tau_{33}) = \frac{1}{2}\big(b(\tau_3)^2 - b(\tau_6)\big)$. The term $b(\tau_3)$ vanishes by (3.21) and $b(\tau_6) = 0$ is a consequence of the proof of Theorem 3.3. Therefore (3.26) is equivalent to $a(\tau_{313}) = 0$. We summarize our computation in the following theorem.

Theorem 3.7. *The processing method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is of order p ($p \leq 8$), if*

- *the coefficients γ_i of the kernel satisfy the conditions of the left column in Example 3.5, i.e., 3 conditions for order 6, and 5 conditions for order 8;*
- *the coefficients δ_i of the processor are such that (3.21) holds (4 conditions for order 6, and 8 conditions for order 8), and in addition condition (3.24) for order 6, and (3.24), (3.25), (3.27) for order 8 are satisfied.* □

Remark 3.8. Although we have presented the computations only for $p \leq 8$, the result is general. All trees $\tau \in \mathcal{H}$, which are not of the form $\tau = u \circ ①$, give rise to conditions on the kernel K_h (for a similar result in the context of Runge–Kutta methods see Butcher & Sanz-Serna (1996)). The remaining conditions have to be satisfied by the coefficients of the processor. Due to the reduced number of order conditions, it is relatively easy to construct high order kernels. However, the difficulty in constructing a suitable processor increases rapidly with the order.

The application of the processing technique is two-fold. A first possibility is to take one of the high-order composition methods of the form (3.2), e.g., one of those presented in Sect. V.3.2, and to exploit the freedom in the coefficients of the processor to make the error constants smaller.

Another possibility is to start from the beginning and to construct a method K_h with coefficients satisfying only the conditions of Theorem 3.7. Methods of effective order 6 and 8 have been constructed in this way by Blanes (2001).

V.4 Symmetric Methods on Manifolds

Numerical methods for differential equations on manifolds have been introduced in Sections IV.4 and IV.5. The presented algorithms are in general not symmetric. We discuss here suitable symmetric modifications which often have an improved long-time behaviour. We consider a differential equation

$$\dot{y} = f(y), \qquad f(y) \in T_y\mathcal{M} \tag{4.1}$$

on a manifold $\mathcal{M}$, and we assume that the manifold is either given as the zero set of a function $g(y)$ or by means of a suitable parametrization $y = \varphi(z)$.

V.4.1 Symmetric Projection

Due to the projection at the end of an integration step, the standard projection method (Algorithm IV.4.2) is not symmetric (see Fig. IV.4.2). In order to make the

overall algorithm symmetric, one has to apply a kind of "inverse projection" at the beginning of each integration step. This idea has first been used by Ascher & Reich (1999) to enforce conservation of energy, and it has been applied in more general contexts by Hairer (2000).

Algorithm 4.1 (Symmetric Projection Method). *Assume that $y_n \in \mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.1, right picture):*

- $\widetilde{y}_n = y_n + G(y_n)^T \mu$ *where $g(y_n) = 0$ (perturbation step);*
- $\widetilde{y}_{n+1} = \Phi_h(\widetilde{y}_n)$ *(symmetric one-step method applied to $\dot{y} = f(y)$);*
- $y_{n+1} = \widetilde{y}_{n+1} + G(y_{n+1})^T \mu$ *with μ such that $g(y_{n+1}) = 0$ (projection step).*

Here, $G(y) = g'(y)$ denotes the Jacobian of $g(y)$. It is important to take a symmetric method in the second step, and the same vector μ in the perturbation and projection steps.

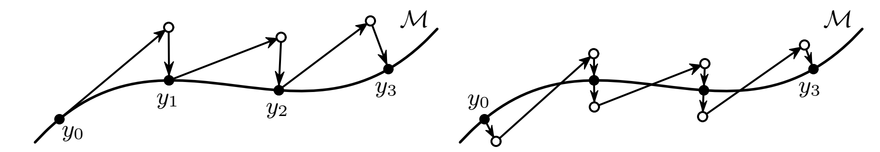


Fig. 4.1. Standard projection (left picture) compared to symmetric projection (right)

Existence of the Numerical Solution. The vector μ and the numerical approximation y_{n+1} are implicitly defined by

$$F(h, y_{n+1}, \mu) = \begin{pmatrix} y_{n+1} - \Phi_h\big(y_n + G(y_n)^T \mu\big) - G(y_{n+1})^T \mu \\ g(y_{n+1}) \end{pmatrix} = 0. \tag{4.2}$$

Since $F(0, y_n, 0) = 0$ and since

$$\frac{\partial F}{\partial(y_{n+1}, \mu)}\big(0, y_n, 0\big) = \begin{pmatrix} I & -2G(y_n)^T \\ G(y_n) & 0 \end{pmatrix} \tag{4.3}$$

is invertible (provided that $G(y_n)$ has full rank), an application of the implicit function theorem proves the existence of the numerical solution for sufficiently small step size h. The simple structure of the matrix (4.3) can also be exploited for an efficient solution of the nonlinear system (4.2) using simplified Newton iterations. If the basic method Φ_h is itself implicit, the nonlinear system (4.2) should be solved in tandem with $\widetilde{y}_{n+1} = \Phi_h(\widetilde{y}_n)$.

Order. For a study of the local error we let $y_n := y(t_n)$ be a value on the exact solution $y(t)$ of (4.1). If the basic method Φ_h is of order p, i.e., if $y(t_n + h) - \Phi_h\big(y(t_n)\big) = \mathcal{O}(h^{p+1})$, we have $F\big(h, y(t_{n+1}), 0\big) = \mathcal{O}(h^{p+1})$. Compared to (4.2) the implicit function theorem yields

$$y_{n+1} - y(t_{n+1}) = \mathcal{O}(h^{p+1}) \qquad \text{and} \qquad \mu = \mathcal{O}(h^{p+1}).$$

This proves that the symmetric projection method of Algorithm 4.1 has the same order as the underlying one-step method Φ_h.

Symmetry of the Algorithm. Exchanging $h \leftrightarrow -h$ and $y_n \leftrightarrow y_{n+1}$ in the Algorithm 4.1 yields

$$\begin{aligned} \widetilde{y}_n &= y_{n+1} + G(y_{n+1})^T\mu, \qquad g(y_{n+1}) = 0, \\ \widetilde{y}_{n+1} &= \Phi_{-h}(\widetilde{y}_n), \\ y_n &= \widetilde{y}_{n+1} + G(y_n)^T\mu, \qquad g(y_n) = 0. \end{aligned}$$

The auxiliary variables μ, $\widetilde{y}_n$, and $\widetilde{y}_{n+1}$ can be arbitrarily renamed. If we replace them with $-\mu$, $\widetilde{y}_{n+1}$, and $\widetilde{y}_n$, respectively, we get the formulas of the original algorithm provided that the method Φ_h of the intermediate step is symmetric. This proves the symmetry of the algorithm.

Various modifications of the perturbation and projection steps are possible without destroying the symmetry. For example, one can replace the arguments y_n and y_{n+1} in $G(y)$ with $(y_n + y_{n+1})/2$. It might be advantageous to use a constant direction, i.e., $\widetilde{y}_n = y_n + A^T\mu,\ y_{n+1} = \widetilde{y}_{n+1} + A^T\mu$ with a constant matrix A. In this case the matrix $G(y)A^T$ has to be invertible along the solution in order to guarantee the existence of the numerical solution.

Reversibility. From Theorem 1.5 we know that symmetry alone does not imply the ρ-reversibility of the numerical flow. The method must also satisfy the compatibility condition (1.4). It is straightforward to check that this condition is satisfied if the integrator Φ_h of the intermediate step of Algorithm 4.1 satisfies (1.4) and, in addition,

$$\rho\, G(y)^T = G(\rho y)^T\sigma \tag{4.4}$$

holds with some constant invertible matrix σ. In many interesting situations we have $g(\rho y) = \sigma^{-T}g(y)$ with a suitable σ, so that (4.4) follows by differentiation if $\rho\rho^T = I$. Similarly, when a projection with constant direction $y = \widetilde{y} + A^T\mu$ is applied, the matrix A has to satisfy $\rho\, A^T = A^T\sigma$ for a suitably chosen invertible matrix σ (see the experiment of Example 4.4 below).

Example 4.2. Let us consider the equations of motion of a rigid body as described in Example IV.1.7. They constitute a differential equation on the manifold

$$\mathcal{M} = \{(y_1, y_2, y_3) \,|\, y_1^2 + y_2^2 + y_3^2 = 1\},$$

and it is ρ-reversible with respect to $\rho(y_1, y_2, y_3) = (-y_1, y_2, y_3)$, and also with respect to $\rho(y_1, y_2, y_3) = (y_1, -y_2, y_3)$ and $\rho(y_1, y_2, y_3) = (y_1, y_2, -y_3)$. For a numerical simulation we take $I_1 = 2$, $I_2 = 1$, $I_3 = 2/3$, and the initial value $y_0 = \big(\cos(0.9), 0, \sin(0.9)\big)$. We apply the trapezoidal rule (II.1.2) with the large step size $h = 1$ in three different versions.

The upper picture of Fig. 4.2 shows the result of a direct application of the trapezoidal rule. The numerical solution lies apparently on a closed curve, but it does not

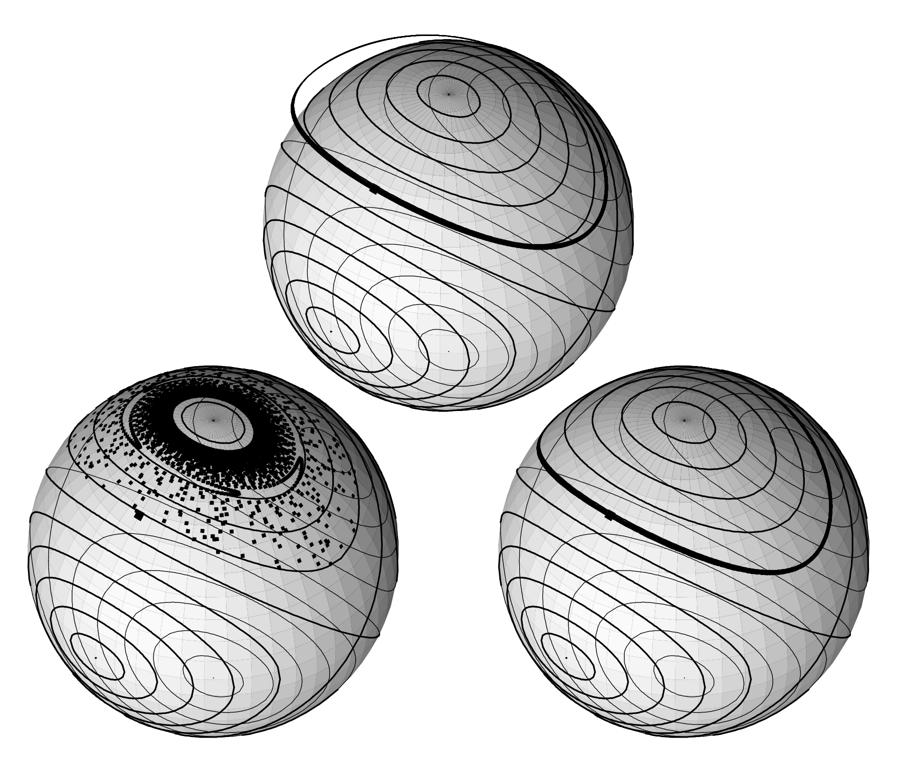


Fig. 4.2. Numerical simulation of the rigid body equations. The three pictures correspond to a direct application (upper), to the standard projection (lower left), and to the symmetric projection (lower right) of the trapezoidal rule; 5000 steps with step size $h = 1$

lie exactly on the manifold $\mathcal{M}$. This can be seen as follows: the trapezoidal rule Φ_h^T is conjugate to the implicit midpoint rule Φ_h^M via a half-step of the explicit Euler method $\chi_{h/2}$. In fact the relations

$$\Phi_h^T = \chi_{h/2}^* \circ \chi_{h/2} \qquad \text{and} \qquad \Phi_h^M = \chi_{h/2} \circ \chi_{h/2}^*$$

hold, so that

$$\Phi_h^T = \chi_{h/2}^{-1} \circ \Phi_h^M \circ \chi_{h/2} \qquad \text{and} \qquad (\Phi_h^T)^N = \chi_{h/2}^{-1} \circ (\Phi_h^M)^N \circ \chi_{h/2}.$$

Consequently, the trajectory of the trapezoidal rule is obtained from the trajectory of the midpoint rule by a simple change of coordinates. On the other hand, the numerical solution of the midpoint rule lies exactly on a solution curve because it conserves quadratic invariants (Theorem IV.2.1).

Using standard orthogonal projection (Algorithm IV.4.2) we obviously obtain a numerical solution lying on the manifold $\mathcal{M}$. But as we can see from the lower left picture of Fig. 4.2, it does not remain near a closed curve and converges to a fixed point. The lower right picture shows that the use of the symmetric orthogonal projection (Algorithm 4.1) recovers the property of remaining near the closed solution curve.

Example 4.3 (Numerical Experiment with Constant Direction of Projection). We consider the pendulum equation in Cartesian coordinates (see Example IV.4.1),

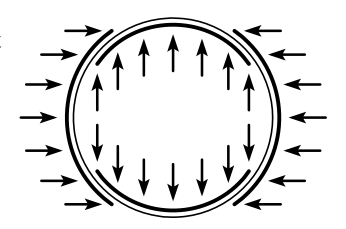
$$\begin{aligned} \dot q_1 &= p_1, & \dot p_1 &= -q_1\lambda, \\ \dot q_2 &= p_2, & \dot p_2 &= -1 - q_2\lambda \end{aligned} \tag{4.5}$$

with $\lambda = (p_1^2 + p_2^2 - q_2)/(q_1^2 + q_2^2)$. This is a problem on the manifold

$$\mathcal{M} = \big\{(q_1, q_2, p_1, p_2) \mid q_1^2 + q_2^2 = 1,\ q_1p_1 + q_2p_2 = 0\big\}.$$

It is ρ-reversible with respect to $\rho(q_1, q_2, p_1, p_2) = (q_1, q_2, -p_1, -p_2)$ and also with respect to $\rho(q_1, q_2, p_1, p_2) = (-q_1, q_2, p_1, -p_2)$.

We apply two kinds of symmetric projection methods. First, we consider an orthogonal projection onto $\mathcal{M}$ as in Algorithm 4.1. Second, we project parallel to coordinate axes. More precisely, we fix the first components in position and velocity if the angle of the pendulum is close to 0 or π (vertical projection in the picture to the right), and we fix the second components if the angle is close to $\pm\pi/2$ (horizontal projection). The regions where the direction of projection changes, are overlapping.



We notice in Fig. 4.3 that for the orthogonal projection method the energy error remains bounded, and this is also true for integrations over much longer time intervals. This is in agreement with the observation of Chap. I, where symmetric methods showed an excellent long-time behaviour when applied to reversible differential equations.

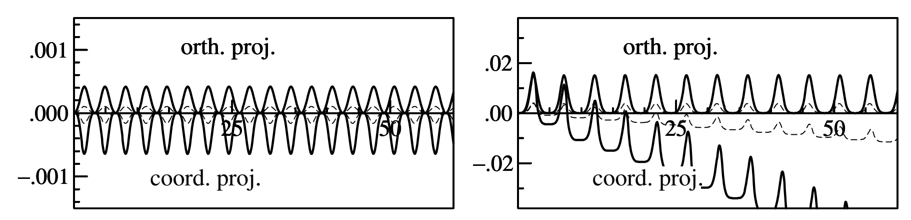


Fig. 4.3. Global error in the total energy for two different projection methods – orthogonal and coordinate projection – with the trapezoidal rule as basic integrator. Initial values for the position are $(\cos 0.8, -\sin 0.8)$ (left picture) and $(\cos 0.8, \sin 0.8)$ (right picture); zero initial values in the velocity; step sizes are $h = 0.1$ (solid) and $h = 0.05$ (thin dashed)

For the coordinate projection, however, we observe a bounded energy error only for the initial value that is close to equilibrium (no change in the direction of the projection is necessary). As soon as the direction has to be changed (right picture of Fig. 4.3) a linear drift in the energy error becomes visible. Hence, care has to be taken with the choice of the projection. For an explanation of this phenomenon we refer to Chap. IX on backward error analysis and to Chap. XI on perturbation theory of reversible mappings.

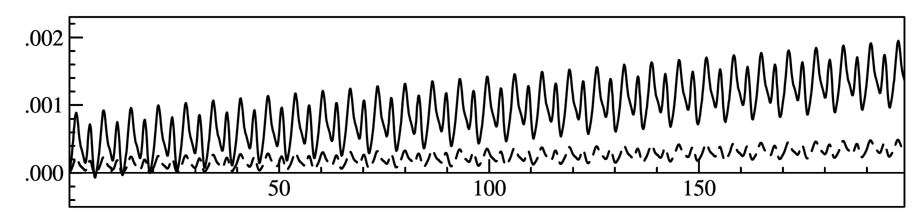


Fig. 4.4. Global error in the total energy for a symmetric projection method violating (1.4). Initial values for the position are $(\cos 0.8, -\sin 0.8)$ and $(0,0)$ for the velocity; step sizes are $h = 0.1$ (solid) and $h = 0.05$ (thin dashed)

Example 4.4 (A Symmetric but Non-Reversible Projection Method). We consider the pendulum equation as in Example 4.3. This time, however, we apply a projection $\widetilde{y}_n = y_n + A^T\mu,\ y_{n+1} = \widetilde{y}_{n+1} + A^T\mu$ with

$$A = \begin{pmatrix} \varepsilon & 1 & 0 & 0 \\ \varepsilon & 0 & 0 & 1 \end{pmatrix}, \qquad \varepsilon = 0.2.$$

For $\varepsilon = 0$ this corresponds to the vertical projection used in Example 4.3. For $\varepsilon \neq 0$ there is no matrix σ such that $\rho A^T = A^T\sigma$ holds for one of the mappings ρ that make the problem ρ-reversible. Hence condition (1.4) is violated, and the method is thus not ρ-reversible. The initial values are chosen such that $g'(y)A^T$ is invertible and well-conditioned along the solution. Although the projection direction need not be changed during the integration and the method is symmetric, the long-time behaviour is disappointing as shown in Fig. 4.4. This experiment illustrates that condition (1.4) is also important for a qualitatively correct simulation.

V.4.2 Symmetric Methods Based on Local Coordinates

Numerical methods for differential equations on manifolds that are based on local coordinates (Algorithm IV.5.3) are in general not symmetric. For example, if we consider the parametrization (IV.5.8) with respect to the tangent space at y_0, the adjoint method would be parametrized by the tangent space at y_1. We can circumvent this difficulty by the following algorithm (Hairer 2001).

Algorithm 4.5 (Symmetric Local Coordinates Approach). *Assume that $y_n \in \mathcal{M}$ and that ψ_a is a local parametrization of $\mathcal{M}$ satisfying $\psi_a(0) = a$ (close to y_n). One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.5):*

- *find z_n (close to 0) such that $\psi_a(z_n) = y_n$;*
- *$\widetilde{z}_{n+1} = \Phi_h(z_n)$ (symmetric one-step method applied to (IV.5.7);*
- *$y_{n+1} = \psi_a(\widetilde{z}_{n+1})$;*
- *choose a in the parametrization such that $z_n + \widetilde{z}_{n+1} = 0$.*

It is important to remark that the parametrization $y = \psi_a(z)$ is in general changed in every step.

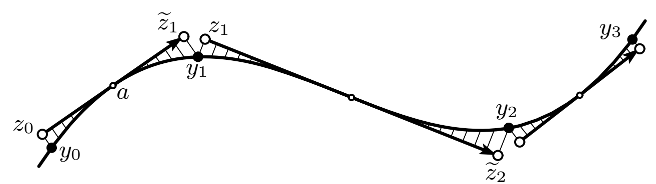


Fig. 4.5. Symmetric use of local tangent space parametrization

This algorithm is illustrated in Fig. 4.5 for the tangent space parametrization (IV.5.8), given by

$$\psi_a(z) = a + Q(a)z + g'(a)^T u_a(z), \tag{4.6}$$

where the columns of $Q(a)$ form an orthogonal basis of $T_a\mathcal{M}$ and the function $u_a(z)$ is such that $\psi_a(z) \in \mathcal{M}$. It satisfies $u_a(0) = 0$ and $u_a'(0) = 0$.

Existence of the Numerical Solution. In Algorithm 4.5 the values $a \in \mathcal{M}$ and z_n are implicitly determined by

$$F(h, z_n, a) = \begin{pmatrix} z_n + \Phi_h(z_n) \\ \psi_a(z_n) - y_n \end{pmatrix} = 0, \tag{4.7}$$

and the numerical solution is then explicitly given by $y_{n+1} = \psi_a\big(\Phi_h(z_n)\big)$. For more clarity we also use here the notation $\psi(z,a) = \psi_a(z)$. If the parametrization $\psi(z,a)$ is differentiable, we have

$$\frac{\partial F}{\partial(z_n, a)}(0, 0, y_n) = \begin{pmatrix} 2I & 0 \\ \frac{\partial \psi}{\partial z}(0, y_n) & \frac{\partial \psi}{\partial a}(0, y_n) \end{pmatrix}. \tag{4.8}$$

Since $\psi(z,a) \in \mathcal{M}$ for all z and $a \in \mathcal{M}$, the derivative with respect to a lies in the tangent space. Assume now that the parametrization $\psi(z,a)$ is such that the restriction of $\frac{\partial \psi}{\partial a}(0, y_n)$ onto the tangent space $T_{y_n}\mathcal{M}$ is bijective. Then, the matrix (4.8) is invertible on $\mathbb{R}^d \times T_{y_n}\mathcal{M}$ (d denotes the dimension of the manifold). The implicit function theorem thus proves the existence of a numerical solution (z_n, a) close to $(0, y_n)$. In the case where $\psi_a(z)$ is given by (4.6), the matrix

$$\frac{\partial \psi}{\partial a}(0, a) = I - g'(a)^T \big(g'(a)g'(a)^T\big)^{-1} g'(a)$$

is a projection onto the tangent space $T_a\mathcal{M}$ and satisfies the above assumptions provided that $g'(a)$ has full rank.

Order. We let $y_n := y(t_n)$ be a value on the exact solution $y(t)$ of (4.1). Then we fix $a \in \mathcal{M}$ as follows: we replace the upper part of the definition (4.7) of $F(h, z_n, a)$ with $z_n + \varphi_h^{(z)}(z_n)$, where $\varphi_t^{(z)}$ denotes the exact flow of the differential equation for $z(t)$ equivalent to (4.1). The above considerations show that such an a exists; let us call it a^*. If Φ_h is of order p, we then have $F\big(h, z(t_n), a^*\big) = \mathcal{O}(h^{p+1})$. An application of the implicit function theorem thus gives $z_n - z(t_n) = \mathcal{O}(h^{p+1})$, implying $\widetilde{z}_{n+1} - z(t_{n+1}) = \mathcal{O}(h^{p+1})$, and finally also $y_{n+1} - y(t_{n+1}) = \mathcal{O}(h^{p+1})$. This proves order p for the method defined by Algorithm 4.5.

Symmetry. Exchanging $h \leftrightarrow -h$ and $y_n \leftrightarrow y_{n+1}$ in Algorithm 4.5 yields

$$\psi_a(z_n) = y_{n+1}, \quad \widetilde{z}_{n+1} = \Phi_{-h}(z_n), \quad y_n = \psi_a(\widetilde{z}_{n+1}), \quad z_n + \widetilde{z}_{n+1} = 0.$$

If we also exchange the auxiliary variables z_n and $\widetilde{z}_{n+1}$ and if we use the symmetry of the basic method Φ_h, we regain the original formulas. This proves the symmetry of the algorithm. Again various kinds of modifications are possible. For example, the condition $z_n + \widetilde{z}_{n+1} = 0$ can be replaced with $z_n + \widetilde{z}_{n+1} = \chi(h, z_n, \widetilde{z}_{n+1})$. If $\chi(-h, v, u) = \chi(h, u, v)$, the symmetry of Algorithm 4.5 is not destroyed.

Reversibility. In general, we cannot expect the method of Algorithm 4.5 to satisfy the ρ-compatibility condition (1.4), which is needed for ρ-reversibility. However, if the parametrization is such that

$$\rho\,\psi_a(z) = \psi_{\rho a}(\sigma z) \qquad \text{for some invertible} \quad \sigma\,, \tag{4.9}$$

we shall show that the compatibility condition (1.4) holds. We first prove that for a ρ-reversible problem $\dot y = f(y)$ the differential equation (IV.5.7), written as $\dot z = F_a(z)$, is σ-reversible in the sense that $\sigma F_a(z) = -F_{\rho a}(\sigma z)$. This follows from $\rho\psi_a'(z) = \psi_{\rho a}'(\sigma z)\sigma$ (which is seen by differentiation of (4.9)) and from $f\big(\psi_{\rho a}(\sigma z)\big) = -\rho\, f\big(\psi_a(z)\big)$, because

$$\psi_a'(z)F_a(z) = f\big(\psi_a(z)\big) \quad \Longrightarrow \quad \psi_{\rho a}'(\sigma z)\sigma F_a(z) = -f\big(\psi_{\rho a}(\sigma z)\big).$$

If the basic method Φ_h satisfies $\sigma \circ \Phi_h = \Phi_{-h} \circ \sigma$ when applied to $\dot z = F_a(z)$ (e.g., for all Runge–Kutta methods), the formulas of Algorithm 4.5 satisfy

$$\begin{aligned}
&\rho\, y_n = \rho\,\psi_a(z_n) = \psi_{\rho a}(\sigma z_n), \qquad \sigma\widetilde{z}_{n+1} = \Phi_{-h}(\sigma z_n),\\
&\psi_{\rho a}(\sigma\widetilde{z}_{n+1}) = \rho\,\psi_a(\widetilde{z}_{n+1}) = \rho\, y_{n+1}, \qquad \sigma z_n + \sigma\widetilde{z}_{n+1} = 0.
\end{aligned}$$

This proves that, starting with ρy_n and a negative step size $-h$, the Algorithm 4.5 produces ρy_{n+1}, where y_{n+1} is just the result obtained with initial value y_n and step size h. But this is nothing other than the ρ-compatibility condition (1.4) for Algorithm 4.5.

In order to verify condition (4.9) for the tangent space parametrization (4.6), we write it as $\psi_a(Z) = a + Z + N(Z)$, where Z is an arbitrary element of the tangent space $T_a\mathcal{M}$ and $N(Z)$ is orthogonal to $T_a\mathcal{M}$ such that $\psi_a(Z) \in \mathcal{M}$. Since $\rho T_a\mathcal{M} = T_{\rho a}\mathcal{M}$ and since, for a ρ satisfying $\rho\rho^T = I$, the vector $\rho N(Z)$ is orthogonal to $T_{\rho a}\mathcal{M}$, we have $\rho\psi_a(Z) = \psi_{\rho a}(\rho Z)$. This proves (4.9) for the tangent space parametrization of a manifold.

Example 4.6. We repeated the experiment of Example 4.2 with Algorithm IV.5.3, using tangent space parametrization and the trapezoidal rule as basic integrator, and compared it to the symmetrized version of Algorithm 4.5. We were surprised to see that both algorithms worked equally well and gave a numerical solution lying near a closed curve. An explanation is given in Exercise 11. There it is shown that for the

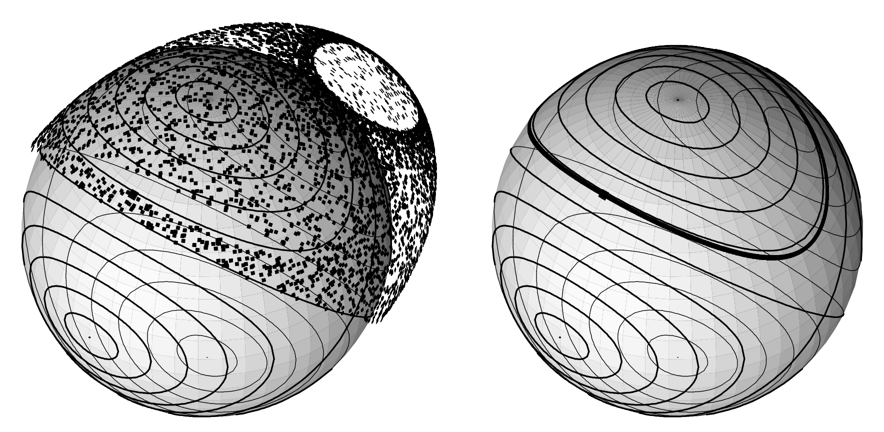


Fig. 4.6. Numerical simulation of the rigid body equations; standard use of tangent space parametrization with the trapezoidal rule as basic method (left picture) and its symmetrized version (right picture); 5000 steps with step size $h = 0.4$

special situation where $\mathcal{M}$ is a sphere, the standard algorithm is also symmetric for the trapezoidal rule. Let us therefore modify the problem slightly.

We consider the rigid body equations (IV.1.4) as a differential equation on the manifold

$$\mathcal{M} = \Big\{ (y_1, y_2, y_3) \,\Big|\, \frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} = Const \Big\} \tag{4.10}$$

with parameters and initial data as in Example 4.2, and we apply the standard and the symmetrized method based on tangent space parametrization. The result is shown in Fig. 4.6. In both cases the numerical solution lies on the manifold (by definition of the method), but only the symmetric method has a correct long-time behaviour.

Symmetric Lie Group Methods. We turn our attention to particular problems

$$\dot{Y} = A(Y)Y, \qquad Y(0) = Y_0, \tag{4.11}$$

where $A(Y)$ is in the Lie algebra $\mathfrak{g}$ whenever Y is in the corresponding Lie group G. The exact solution then evolves on the manifold G. Munthe-Kaas methods (Sect. IV.8.2) are in general not symmetric, even if the underlying Runge–Kutta method is symmetric. This is due to the unsymmetric use of the local coordinates $Y = \exp(\Omega)Y_0$. However, accidentally, the Lie group method based on the implicit midpoint rule

$$Y_{n+1} = \exp(\Omega)Y_n, \qquad \Omega = hA\big(\exp(\Omega/2)Y_n\big) \tag{4.12}$$

is symmetric. This can be seen as usual by exchanging $h \leftrightarrow -h$ and $Y_n \leftrightarrow Y_{n+1}$ (and also $\Omega \leftrightarrow -\Omega$ for the auxiliary variable). Numerical computations with the rigid body equations (considered as a problem on the sphere) shows an excellent long-time behaviour for this method similar to that of the right picture in Fig. 4.6. In contrast to the implicit midpoint rule (I.1.7), the numerical solution of (4.12) does not lie exactly on the ellipsoid (4.10); see Exercise 12.

For the construction of further symmetric Lie group methods we can apply the ideas of Algorithm 4.5. As local parametrization we choose

$$\psi_U(\Omega) = \exp(\Omega)U, \tag{4.13}$$

where $U = \exp(\Theta)Y_n$ plays the role of the midpoint on the manifold. We put $Z_n = -\Theta$ so that $\psi_U(Z_n) = Y_n$. With this starting value Z_n we apply any symmetric Runge–Kutta method to the differential equation

$$\dot{\Omega} = A\big(\psi_U(\Omega)\big) + \sum_{k=1}^{q} \frac{B_k}{k!}\,\mathrm{ad}_{\Omega}^{\,k}\Big(A\big(\psi_U(\Omega)\big)\Big), \qquad \Omega(0) = -\Theta, \tag{4.14}$$

(cf. (IV.8.9)) and thus obtain $\widetilde{Z}_{n+1}$. According to Algorithm 4.5, Θ is implicitly determined by the condition $Z_n + \widetilde{Z}_{n+1} = 0$, and the numerical approximation is obtained from

$$Y_{n+1} = \psi_U(\widetilde{Z}_{n+1}) = \exp(\widetilde{Z}_{n+1})\exp(\Theta)Y_n = \exp(2\Theta)Y_n.$$

The method obtained in this way is identical to Algorithm 2 of Zanna, Engø & Munthe-Kaas (2001). With the coefficients of the 2-stage Gauss method (Table II.1.1) and with $q = 1$ in (4.14) we thus get

$$\begin{aligned} \Omega_1 &= -h\frac{\sqrt{3}}{6}\Big(A_2 - \frac{1}{2}[\Omega_2, A_2]\Big), \qquad \Omega_2 = h\frac{\sqrt{3}}{6}\Big(A_1 - \frac{1}{2}[\Omega_1, A_1]\Big) \\ Y_{n+1} &= \exp\big(2\Theta\big)Y_n = \exp\Big(\frac{h}{2}\big(A_1 + A_2\big) - \frac{h}{4}\big([\Omega_1, A_1] + [\Omega_2, A_2]\big)\Big)Y_n, \end{aligned}$$

where $A_i = A\big(\exp(\Omega_i)\exp(\Theta)Y_n\big)$. This is a symmetric Lie group method of order four. We can reduce the number of commutators by replacing Ω_i in the right-hand expression with its dominating term. This yields

$$\begin{aligned} \Omega_1 &= -h\frac{\sqrt{3}}{6}A_2 + \frac{h^2}{24}[A_1, A_2], \qquad \Omega_2 = h\frac{\sqrt{3}}{6}A_1 - \frac{h^2}{24}[A_1, A_2] \\ Y_{n+1} &= \exp\Big(\frac{h}{2}\big(A_1 + A_2\big) - h^2\frac{\sqrt{3}}{12}[A_1, A_2]\Big)Y_n \end{aligned} \tag{4.15}$$

(cf. Exercise IV.19). Although we have neglected terms of size $\mathcal{O}(h^4)$, the method remains of order four, because the order of symmetric methods is always even.

For any linear invertible transformation ρ, the parametrization (4.13) satisfies

$$\rho\,\psi_U(\Omega) = \rho\,\exp(\Omega)U = \exp(\rho\Omega\rho^{-1})\rho U = \psi_{\rho U}(\sigma U)$$

with $\sigma\Omega = \rho\Omega\rho^{-1}$. Hence (4.9) holds true. If the problem (4.11) is ρ-reversible, i.e., $\rho A(Y) = -A(\rho Y)\rho$, then the truncated differential equation (4.14) is σ-reversible for all choices of the truncation index q. Moreover, after the simplifications that lead to method (4.15), the ρ-compatibility condition (1.4) is also satisfied.

The following variant is also proposed in Zanna, Engø & Munthe-Kaas (2001). Instead of computing Θ from the relation $Z_n + \widetilde{Z}_{n+1} = 0$, Θ is determined by

$$Z_n + \widetilde{Z}_{n+1} = h \sum_{i=1}^{s} e_i \Big(A_i - \frac{1}{2}[\Omega_i, A_i] + \ldots \Big).$$

If the coefficients satisfy $e_{s+1-i} = -e_i$, this modification gives symmetric Lie group methods.

V.5 Energy – Momentum Methods and Discrete Gradients

> Conventional numerical methods, when applied to the ordinary differential equations of motion of classical mechanics, conserve the total energy and angular momentum only to the order of the truncation error. Since these constants of motion play a central role in mechanics, it is a great advantage to be able to conserve them exactly.
>
> (R.A. LaBudde & D. Greenspan 1976)

This section is concerned with numerical integrators for the equations of motion of classical mechanics which conserve both the total energy and angular momentum. Their construction is related to the concept of discrete gradients. The methods considered are symmetric, which is incidental but useful: in our view their good long-time behaviour is a consequence of their symmetry (and reversibility) more than of their exact conservation properties; see the disappointing behaviour of the non-symmetric energy- and momentum-conserving projection method in Example IV.4.4.

A Modified Midpoint Rule. Consider first a single particle of mass m in $\mathbb{R}^3$, with position coordinates $q(t) \in \mathbb{R}^3$, moving in a central force field with potential $U(q) = V(\|q\|)$ (e.g., $V(r) = -1/r$ in the Kepler problem). With the momenta $p(t) = m\,\dot{q}(t)$, the equations of motion read

$$\dot{q} = \frac{1}{m}\, p\,, \qquad \dot{p} = -\nabla U(q) = -V'(\|q\|)\, \frac{q}{\|q\|}\,.$$

Constants of motion are the total energy $H = T(p) + U(q)$, with $T(p) = \|p\|^2/(2m)$, and the angular momentum $L = q \times p$:

$$\frac{d}{dt}(q \times p) = \dot{q} \times p + q \times \dot{p} = \frac{1}{m}\, p \times p - V'(\|q\|)\, \frac{1}{\|q\|}\, q \times q = 0\,.$$

We know from Sect. IV.2 that the implicit midpoint rule conserves the quadratic invariant $L = q \times p$, and Theorem IV.2.4 (or a simple direct calculation) shows that L remains actually conserved by any modification of the form

$$\begin{aligned} q_{n+1} &= q_n + \frac{h}{m}\, p_{n+1/2} \\ p_{n+1} &= p_n - \kappa h \nabla U(q_{n+1/2}) \end{aligned} \qquad \text{with} \qquad \begin{aligned} p_{n+1/2} &= \tfrac{1}{2}(p_n + p_{n+1}) \\ q_{n+1/2} &= \tfrac{1}{2}(q_n + q_{n+1}) \end{aligned} \tag{5.1}$$

where κ is an arbitrary real number. Simo, Tarnow & Wong (1992) introduce this additional parameter κ and determine it so that the total energy is conserved: $H(p_{n+1}, q_{n+1}) = H(p_n, q_n)$. With the notation $F_{n+1/2} = -\nabla U(q_{n+1/2}) = -V'(\|q_{n+1/2}\|)/\|q_{n+1/2}\| \cdot q_{n+1/2}$ we have

$$T(p_{n+1}) = T(p_n + \kappa h F_{n+1/2}) = T(p_n) + \frac{\kappa h}{m}\, p_{n+1/2}^T\, F_{n+1/2}\ ,$$

and hence the condition for conservation of the total energy $H = T + U$ becomes

$$\kappa\, \frac{h}{m}\, p_{n+1/2}^T\, F_{n+1/2} = U(q_n) - U(q_{n+1})\ .$$

This gives a reasonable method even if $p_{n+1/2}^T\, F_{n+1/2}$ is arbitrarily close to zero. This is seen as follows: let $\sigma = -\kappa V'(\|q_{n+1/2}\|)/\|q_{n+1/2}\|$ so that $\kappa F_{n+1/2} = \sigma q_{n+1/2}$. The above condition for energy conservation then reads

$$\sigma\, \frac{h}{m}\, p_{n+1/2}^T\, q_{n+1/2} = V(\|q_n\|) - V(\|q_{n+1}\|)\ ,$$

where we note further that

$$\begin{aligned} \frac{h}{m}\, p_{n+1/2}^T\, q_{n+1/2} &= (q_{n+1} - q_n)^T\, \tfrac{1}{2}(q_{n+1} + q_n) \\ &= \tfrac{1}{2}\big(\|q_{n+1}\|^2 - \|q_n\|^2\big) = \big(\|q_{n+1}\| - \|q_n\|\big)\, \tfrac{1}{2}\big(\|q_{n+1}\| + \|q_n\|\big)\ . \end{aligned}$$

These formulas give

$$\sigma = -\frac{V(\|q_{n+1}\|) - V(\|q_n\|)}{\|q_{n+1}\| - \|q_n\|}\, \frac{1}{\tfrac{1}{2}\big(\|q_{n+1}\| + \|q_n\|\big)}\ , \tag{5.2}$$

with which method (5.1) becomes

$$\begin{aligned} q_{n+1} &= q_n + \frac{h}{m}\, p_{n+1/2} \\ p_{n+1} &= p_n - h\, \frac{V(\|q_{n+1}\|) - V(\|q_n\|)}{\|q_{n+1}\| - \|q_n\|}\, \frac{q_{n+1/2}}{\tfrac{1}{2}\big(\|q_{n+1}\| + \|q_n\|\big)}\ . \end{aligned} \tag{5.3}$$

This is a second-order symmetric method which conserves the total energy and the angular momentum. It evaluates only the potential $U(q) = V(\|q\|)$. The force $-\nabla U(q) = -V'(\|q\|)\, \frac{q}{\|q\|}$ is approximated by finite differences.

The energy- and momentum-conserving method (5.3) first appeared in LaBudde & Greenspan (1974). The method (5.1) or (5.3) is the starting point for extensions in several directions to other problems of mechanics and other methods; see Simo,

Tarnow & Wong (1992), Simo & Tarnow (1992), Lewis & Simo (1994, 1996), Gonzalez & Simo (1996), Gonzalez (1996), and Reich (1996b). In the following we consider a direct generalization to systems of particles, also given in LaBudde & Greenspan (1974).

An Energy-Momentum Method for N-Body Systems. We consider a system of N particles interacting pairwise with potential forces which depend on the distances between the particles. As in Example IV.1.3, this is formulated as a Hamiltonian system with total energy

$$H(p,q) = \frac{1}{2}\sum_{i=1}^{N}\frac{1}{m_i}\,p_i^T p_i + \sum_{i=2}^{N}\sum_{j=1}^{i-1} V_{ij}\Big(\|q_i - q_j\|\Big)\,. \tag{5.4}$$

As an extension of method (5.3), we consider the following method (where we now write the time step number as a superscript for notational convenience)

$$\begin{aligned} q_i^{n+1} &= q_i^n + \frac{h}{m_i}\,p_i^{n+1/2} \\ p_i^{n+1} &= p_i^n + h\sum_{j=1}^{N}\sigma_{ij}\left(q_i^{n+1/2} - q_j^{n+1/2}\right) \end{aligned} \tag{5.5}$$

where $p_i^{n+1/2} = \frac{1}{2}(p_i^n + p_i^{n+1})$, $q_i^{n+1/2} = \frac{1}{2}(q_i^n + q_i^{n+1})$, and for $i > j$,

$$\sigma_{ij} = \sigma_{ji} = -\frac{V_{ij}(r_{ij}^{n+1}) - V_{ij}(r_{ij}^n)}{r_{ij}^{n+1} - r_{ij}^n}\,\frac{1}{\frac{1}{2}(r_{ij}^n + r_{ij}^{n+1})} \tag{5.6}$$

with $r_{ij}^n = \|q_i^n - q_j^n\|$, and $\sigma_{ii} = 0$. This method has the following properties.

Theorem 5.1 (LaBudde & Greenspan 1974). *The method (5.5) with (5.6) is a second-order symmetric implicit method which conserves the total linear momentum $P = \sum_{i=1}^N p_i$, the total angular momentum $L = \sum_{i=1}^N q_i \times p_i$, and the total energy H.*

Proof. A comparison of (5.6) with the equations of motion shows that the method is of order 2. Similar to the continuous case (Example IV.1.3), the conservation of linear and angular momentum is obtained as a consequence of the symmetry $\sigma_{ij} = \sigma_{ji}$ for all i, j. For the linear momentum we have

$$\sum_{i=1}^{N} p_i^{n+1} = \sum_{i=1}^{N} p_i^n + h\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{ij}\,(q_i^{n+1/2} - q_j^{n+1/2}) = \sum_{i=1}^{N} p_i^n\,.$$

For the proof of the conservation of the angular momentum we observe that the first equation of (5.5) together with $p_i^{n+1/2} = \frac{1}{2}(p_i^{n+1} + p_i^n)$ yields

$$\left(q_i^{n+1} - q_i^n\right) \times \left(p_i^{n+1} + p_i^n\right) \;=\; 0 \tag{5.7}$$

for all i. The second equation of (5.5) together with $q_i^{n+1/2} = \frac{1}{2}(q_i^{n+1} + q_i^n)$ gives

$$\sum_{i=1}^{N} \big(q_i^{n+1} + q_i^n\big) \times \big(p_i^{n+1} - p_i^n\big) \;=\; 0\,, \tag{5.8}$$

because $\sigma_{ij} = \sigma_{ji}$ and therefore $\sum_{i,j=1}^{N} \sigma_{ij}\, q_i^{n+1/2} \times q_j^{n+1/2} = 0$. Adding the sum over i of (5.7) to the equation (5.8) proves the statement $\sum_{i=1}^{N} q_i^{n+1} \times p_i^{n+1} = \sum_{i=1}^{N} q_i^n \times p_i^n$.

It remains to show the energy conservation. Now, the kinetic energy $T(p) = \frac{1}{2}\sum_{i=1}^{N} m_i^{-1} p_i^T p_i$ at step $n+1$ is

$$\begin{aligned} T(p^{n+1}) \;&=\; T(p^n) + \sum_{i=1}^{N} \big(\frac{h}{m_i}\, p_i^{n+1/2}\big)^T \sum_{j=1}^{N} \sigma_{ij}\,\big(q_i^{n+1/2} - q_j^{n+1/2}\big) \\ &=\; T(p^n) + \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_{ij}\,\big(q_i^{n+1} - q_i^n\big)^T \big(q_i^{n+1/2} - q_j^{n+1/2}\big)\,. \end{aligned}$$

Using once more the symmetry $\sigma_{ij} = \sigma_{ji}$, the double sum reduces to

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_{ij}\, \Big(\big(q_i^{n+1} - q_j^{n+1}\big) - \big(q_i^n - q_j^n\big)\Big)^T \frac{1}{2}\Big(\big(q_i^{n+1} - q_j^{n+1}\big) + \big(q_i^n - q_j^n\big)\Big) \\ &= \sum_{i=2}^{N}\sum_{j=1}^{i-1} \sigma_{ij}\, \frac{1}{2}\Big(\big(r_{ij}^{n+1}\big)^2 - \big(r_{ij}^n\big)^2\Big)\,. \end{aligned}$$

On the other hand, the change in the potential energy is

$$U(q^{n+1}) - U(q^n) \;=\; \sum_{i=2}^{N}\sum_{j=1}^{i-1}\Big(V_{ij}(r_{ij}^{n+1}) - V_{ij}(r_{ij}^n)\Big)\,,$$

and hence (5.6) yields the conservation of the total energy $H = T + U$. □

Discrete-Gradient Methods. The methods (5.3) and (5.5) are of the form

$$y_{n+1} \;=\; y_n + h\overline{B}(y_{n+1}, y_n)\,\overline{\nabla} H(y_{n+1}, y_n) \tag{5.9}$$

where $\overline{B}(\widehat{y}, y)$ is a skew-symmetric matrix for all $\widehat{y}, y$, and $\overline{\nabla} H(\widehat{y}, y)$ is a *discrete gradient* of H, that is, a continuous function of $(\widehat{y}, y)$ satisfying

$$\begin{aligned} \overline{\nabla} H(\widehat{y}, y)^T(\widehat{y} - y) \;&=\; H(\widehat{y}) - H(y) \\ \overline{\nabla} H(y, y) \;&=\; \nabla H(y)\,. \end{aligned} \tag{5.10}$$

The symmetry of the methods is seen from the properties $\overline{B}(\widehat{y}, y) = \overline{B}(y, \widehat{y})$ and $\overline{\nabla} H(\widehat{y}, y) = \overline{\nabla} H(y, \widehat{y})$. For example, for method (5.3) we have, with $y = (p, q)$ and $\widehat{y} = (\widehat{p}, \widehat{q})$,

$$\overline{B}(\widehat{y}, y) = \begin{pmatrix} 0 & -I_3 \\ I_3 & 0 \end{pmatrix} \quad \text{and} \quad \overline{\nabla} H(\widehat{y}, y) = \begin{pmatrix} \frac{1}{2}(\widehat{p} + p) \\ \sigma(\widehat{q}, q)\, \frac{1}{2}(\widehat{q} + q) \end{pmatrix}$$

where $\sigma(\widehat{q}, q)$ is given by the expression (5.2) with $(\widehat{q}, q)$ in place of (q_{n+1}, q_n) or by the corresponding limit as $\|\widehat{q}\| \to \|q\|$.

The discrete-gradient method (5.9) is consistent with the differential equation

$$\dot{y} = B(y)\, \nabla H(y) \tag{5.11}$$

with the skew-symmetric matrix $B(y) = \overline{B}(y, y)$. This system conserves H, since

$$\frac{d}{dt}\, H(y) = \nabla H(y)^T \dot{y} = \nabla H(y)^T\, B(y)\, \nabla H(y) = 0\,,$$

and, as was noted by Gonzalez (1996) and McLachlan, Quispel & Robidoux (1999), H is also conserved by method (5.9).

Theorem 5.2. *The discrete-gradient method (5.9) conserves the invariant H of the system (5.11).*

Proof. The definitions (5.10) of a discrete gradient and of the method (5.9) give

$$\begin{aligned} H(y_{n+1}) - H(y_n) &= \overline{\nabla} H(y_{n+1}, y_n)^T (y_{n+1} - y_n) \\ &= h\, \overline{\nabla} H(y_{n+1}, y_n)^T\, \overline{B}(y_{n+1}, y_n)\, \overline{\nabla} H(y_{n+1}, y_n) = 0\,, \end{aligned}$$

where the last equality follows from the skew-symmetry of $\overline{B}(y_{n+1}, y_n)$. □

Example 5.3. The Lotka–Volterra system (I.1.1) can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & -uv \\ uv & 0 \end{pmatrix} \nabla H(u, v)$$

with the invariant $H(u, v) = \ln u - u + 2 \ln v - v$ of (I.1.4). Possible choices of a discrete gradient are the *coordinate increment discrete gradient* (Itoh & Abe 1988)

$$\overline{\nabla} H(\widehat{u}, \widehat{v}; u, v) = \begin{pmatrix} \dfrac{H(\widehat{u}, v) - H(u, v)}{\widehat{u} - u} \\ \dfrac{H(\widehat{u}, \widehat{v}) - H(\widehat{u}, v)}{\widehat{v} - v} \end{pmatrix} \tag{5.12}$$

and the *midpoint discrete gradient* (Gonzalez 1996)

$$\overline{\nabla} H(\widehat{y}, y) \;=\; \nabla H(\overline{y}) + \frac{H(\widehat{y}) - H(y) - \nabla H(\overline{y})^T \Delta y}{\|\Delta y\|^2}\, \Delta y \tag{5.13}$$

with $\overline{y} = \frac{1}{2}(\widehat{y} + y)$ and $\Delta y = \widehat{y} - y$. In contrast to (5.12), the discrete gradient (5.13) yields a symmetric discretization.

A systematic study of discrete-gradient methods is given in Gonzalez (1996) and McLachlan, Quispel & Robidoux (1999).

V.6 Exercises

1. Prove that (after a suitable permutation of the stages) the condition $c_{s+1-i} = 1 - c_i$ (for all i) is also necessary for a collocation method to be symmetric.
2. Prove that explicit Runge–Kutta methods cannot be symmetric.
   Hint. If a one-step method applied to $\dot{y} = \lambda y$ yields $y_1 = R(h\lambda)y_0$ then, a necessary condition for the symmetry of the method is $R(z)R(-z) = 1$ for all complex z.
3. Consider an irreducible diagonally implicit Runge–Kutta method (irreducible in the sense of Sect. VI.7.3). Prove that the condition (2.4) is necessary for the symmetry of the method. No permutation of the stages has to be performed.
4. Let $\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$, where $\varphi_t^{[i]}$ represents the exact flow of $\dot{y} = f^{[i]}(y)$. In the situation of Theorem III.3.17 prove that the local error (3.4) of the composition method (3.3) has the form
$$h^3\Big(\frac{1}{24}(6\alpha - 1)\big[D_2, [D_2, D_1]\big] + \frac{1}{12}(1 - 6\alpha + 6\alpha^2)\big[D_1, [D_1, D_2]\big]\Big) Id(y),$$
   where, as usual, $D_i g(y) = g'(y) f^{[i]}(y)$. The value $\alpha = 0.1932$ is found by minimizing the expression $(6\alpha - 1)^2 + 4(1 - 6\alpha + 6\alpha^2)^2$ (McLachlan 1995).
5. For the linear transformation $\rho(p, q) = (-p, q)$, consider a ρ-reversible problem (1.3) with scalar p and q. Prove that every solution which crosses the q-axis twice is periodic.
6. Prove that if a numerical method conserves quadratic invariants (IV.2.1), then so does its adjoint.
7. For the numerical solution of $\dot{y} = A(t)y$ consider the method $y_n \mapsto y_{n+1}$ defined by $y_{n+1} = z(t_n + h)$, where $z(t)$ is the solution of
$$\dot{z} = \widehat{A}(t)z, \qquad z(t_n) = y_n,$$
   and $\widehat{A}(t)$ is the interpolation polynomial based on symmetric nodes $c_1, \ldots, c_s$, i.e., $c_{s+1-i} + c_i = 1$ for all i.
   a) Prove that this method is symmetric.
   b) Show that $y_{n+1} = \exp(\Omega(h))y_n$ holds, where $\Omega(h)$ has an expansion in odd powers of h. This justifies the omission of the terms involving triple integrals in Example IV.7.4.
8. If Φ_h stands for the implicit midpoint rule, what are the Runge–Kutta coefficients of the composition method (3.8)? The general theory of Sect. III.1 gives three order conditions for order 4 (those for the trees of order 2 and 4 are automatically satisfied by the symmetry of the method). Are they compatible with the two conditions of Example 3.5?
9. Make a numerical comparison of our favourite composition methods $p6\,s9$, $p8\,s17$, and $p10\,s35$ for the Lorenz problem
$$\begin{array}{lll} y_1' = -\sigma(y_1 - y_2) & y_1(0) = 10 & \sigma = 10 \\ y_2' = -y_1 y_3 + r y_1 - y_2 & y_2(0) = -20 & r = 28 \\ y_3' = y_1 y_2 - b y_3 & y_3(0) = 20 & b = 8/3 \end{array} \tag{6.1}$$

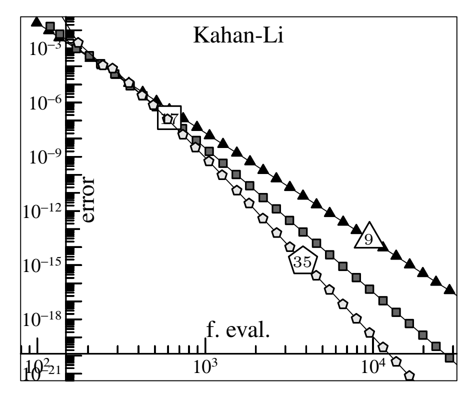


Fig. 6.1. Comparison of various composition methods applied to the Lorenz equations

with exact solution

$$\begin{aligned} y_1(1) &= 8.6356927098925060179305446 28639 \\ y_2(1) &= 2.7986633879274570520230800 59065 \\ y_3(1) &= 33.360635089731421577891858 46267 \end{aligned} \tag{6.2}$$

by composing for $0 \le t \le 1$ the second order *symmetric splitting* scheme (see Kahan & Li 1997) which, for the time-stepping $y_i \mapsto Y_i$, is given by

$$\begin{aligned} Y_1 - y_1 &= \frac{h}{2}\big(-\sigma(y_1 + Y_1 - y_2 - Y_2)\big) \\ Y_2 - y_2 &= \frac{h}{2}\big(-y_1 Y_3 - Y_1 y_3 + r y_1 + r Y_1 - y_2 - Y_2\big) \\ Y_3 - y_3 &= \frac{h}{2}\big(y_1 Y_2 + Y_1 y_2 - b y_3 - b Y_3\big). \end{aligned} \tag{6.3}$$

This method requires, for each step, the solution of a *linear* system only. The results are shown in Fig. 6.1.

10. *Symmetrized order conditions* (Suzuki 1992). Prove that for methods (3.8) of order four with γ_i satisfying (3.10)

$$\sum_{k=1}^{s} \gamma_k^3 \Big(\sum_{\ell=1}^{k}{}' \gamma_\ell \Big)^2 = 0 \qquad \Longleftrightarrow \qquad \sum_{k=1}^{s} \gamma_k^3 \Big(\sum_{\ell=1}^{k}{}' \gamma_\ell \Big) \Big({}' \sum_{\ell=k}^{s} \gamma_\ell \Big) = 0.$$

The prime after (before) a sum sign indicates that the term with highest (lowest) index is divided by 2. Prove also that the order conditions given in Suzuki (1992) for order $p \le 8$ are equivalent to those of Example 3.5. Is this also true for order $p = 10$?

Hint. Use relations like $\sum_{\ell=1}^{k}{}' \gamma_\ell = 1 - {}'\sum_{\ell=k}^{s} \gamma_\ell$.

11. Let $\mathcal{M} = \big\{ (y_1, y_2, y_3) \,|\, y_1^2 + y_2^2 + y_3^2 = 1 \big\}$, and consider for $a \in \mathcal{M}$ the tangent space parametrization

$$\psi_a(z) = a + z + a u_a(z),$$

where, for $z \in T_a\mathcal{M}$, the real value $u_a(z)$ is determined by the requirement $\psi_a(z) \in \mathcal{M}$. Prove that Algorithm IV.5.3, with the trapezoidal rule in the role of Φ_h, is a symmetric method.
Hint. Since z is a linear combination of a and $\psi_a(z)$, it is uniquely determined by $a^T z$ (which is zero) and $\psi_a(z)^T z$.

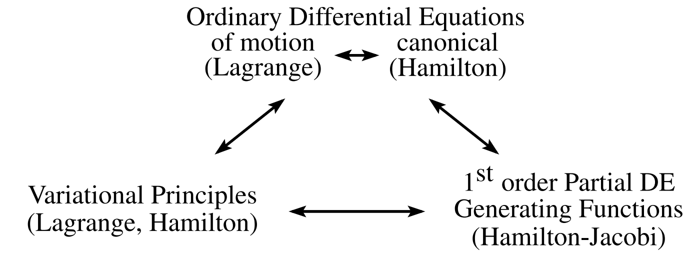
12. (Zanna, Engø & Munthe-Kaas 2001). Verify numerically that the Lie group method (4.12) based on the implicit midpoint rule does not conserve general quadratic first integrals. One can consider the rigid body equations in the form (IV.1.5).

Chapter VI.
Symplectic Integration of Hamiltonian Systems



Fig. 0.1. Sir William Rowan Hamilton, born: 4 August 1805 in Dublin, died: 2 September 1865. Famous for research in optics, mechanics, and for the invention of quaternions

Hamiltonian systems form the most important class of ordinary differential equations in the context of 'Geometric Numerical Integration'. An outstanding property of these systems is the symplecticity of the flow. As indicated in the following diagram,



Hamiltonian theory operates in three different domains (equations of motion, partial differential equations and variational principles) which are all interconnected. Each of these viewpoints, which we will study one after the other, leads to the construction of methods preserving the symplecticity.

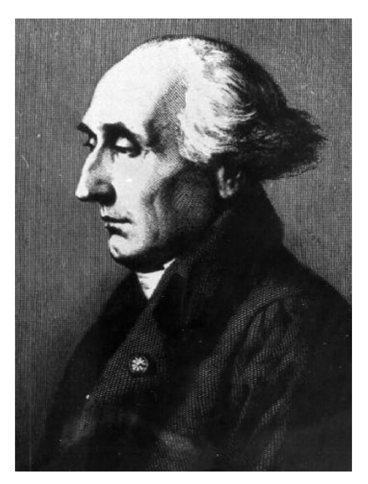
VI.1 Hamiltonian Systems

Hamilton's equations appeared first, among thousands of other formulas, and inspired by previous research in optics, in Hamilton (1834). Their importance was immediately recognized by Jacobi, who stressed and extended the fundamental ideas, so that, a couple of years later, all the long history of research of Galilei, Newton, Euler and Lagrange, was, in the words of Jacobi (1842), "to be considered as an introduction". The next mile-stones in the exposition of the theory were the monumental three volumes of Poincaré (1892,1893,1899) on celestial mechanics, Siegel's "Lectures on Celestial Mechanics" (1956), English enlarged edition by Siegel & Moser (1971), and the influential book of V.I. Arnold (1989; first Russian edition 1974). Beyond that, Hamiltonian systems became fundamental in many branches of physics. One such area, the dynamics of particle accelerators, actually motivated the construction of the first symplectic integrators (Ruth 1983).

VI.1.1 Lagrange's Equations

> Équations différentielles pour la solution de tous les problèmes de Dynamique. (J.-L. Lagrange 1788)

The problem of computing the dynamics of general mechanical systems began with Galilei (published 1638) and Newton's *Principia* (1687). The latter allowed one to reduce the movement of free mass points (the "mass points" being such planets as Mars or Jupiter) to the solution of differential equations (see Sect. I.2). But the movement of more complicated systems such as rigid bodies or bodies attached to each other by rods or springs, were the subject of long and difficult developments, until Lagrange (1760, 1788) found an elegant way of treating such problems in general.



Joseph-Louis Lagrange[1]

We suppose that the position of a mechanical system with d degrees of freedom is described by $q = (q_1, \ldots, q_d)^T$ as *generalized coordinates* (this can be for example Cartesian coordinates, angles, arc lengths along a curve, etc.). The theory is then built upon two pillars, namely an expression

$$T = T(q, \dot{q}) \tag{1.1}$$

which represents the *kinetic energy* (and which is often of the form $\frac{1}{2}\dot{q}^T M(q)\dot{q}$ where $M(q)$ is symmetric and positive definite), and by a function

[1] Joseph-Louis Lagrange, born: 25 January 1736 in Turin, Sardinia–Piedmont (now Italy), died: 10 April 1813 in Paris.

$$U = U(q) \tag{1.2}$$

representing the *potential energy*. Then, after denoting by

$$L = T - U \tag{1.3}$$

the corresponding *Lagrangian*, the coordinates $q_1(t), \ldots, q_d(t)$ obey the differential equations

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot q}\Big) = \frac{\partial L}{\partial q}\ , \tag{1.4}$$

which constitute the *Lagrange equations* of the system. A numerical (or analytical) integration of these equations allows one to predict the motion of any such system from given initial values ("Ce sont ces équations qui serviront à déterminer la courbe décrite par le corps M et sa vitesse à chaque instant"; Lagrange 1760, p. 369).

Example 1.1. For a mass point of mass m in $\mathbb{R}^3$ with Cartesian coordinates $x = (x_1, x_2, x_3)^T$ we have $T(\dot x) = m(\dot x_1^2 + \dot x_2^2 + \dot x_3^2)/2$. We suppose the point to move in a conservative force field $F(x) = -\nabla U(x)$. Then, the Lagrange equations (1.4) become $m\ddot x = F(x)$, which is Newton's second law. The equations (I.2.2) for the planetary motion are precisely of this form.

Example 1.2 (Pendulum). For the mathematical pendulum of Sect. I.1 we take the angle α as coordinate. The kinetic and potential energies are given by $T = m(\dot x^2 + \dot y^2)/2 = m\ell^2\dot\alpha^2/2$ and $U = mgy = -mg\ell\cos\alpha$, respectively, so that the Lagrange equations become $-mg\ell\sin\alpha - m\ell^2\ddot\alpha = 0$ or equivalently $\ddot\alpha + \frac{g}{\ell}\sin\alpha = 0$.

VI.1.2 Hamilton's Canonical Equations

> An diese *Hamiltonsche* Form der Differentialgleichungen werden die ferneren Untersuchungen, welche den Kern dieser Vorlesung bilden, anknüpfen; das Bisherige ist als Einleitung dazu anzusehen.
> (C.G.J. Jacobi 1842, p. 143)

Hamilton (1834) simplified the structure of Lagrange's equations and turned them into a form that has remarkable symmetry, by

- introducing Poisson's variables, the conjugate *momenta*

$$p_k = \frac{\partial L}{\partial \dot q_k}(q, \dot q) \qquad \text{for} \quad k = 1, \ldots, d, \tag{1.5}$$

- considering the *Hamiltonian*

$$H := p^T \dot q - L(q, \dot q) \tag{1.6}$$

  as a function of p and q, i.e., taking $H = H(p, q)$ obtained by expressing $\dot q$ as a function of p and q via (1.5).

Here it is, of course, required that (1.5) defines, for every q, a continuously differentiable bijection $\dot q \leftrightarrow p$. This map is called the *Legendre transform*.

Theorem 1.3. *Lagrange's equations (1.4) are equivalent to Hamilton's equations*

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}(p,q), \qquad \dot{q}_k = \frac{\partial H}{\partial p_k}(p,q), \qquad k = 1,\ldots,d. \tag{1.7}$$

Proof. The definitions (1.5) and (1.6) for the momenta p and for the Hamiltonian H imply that

$$\begin{aligned} \frac{\partial H}{\partial p} &= \dot{q}^T + p^T \frac{\partial \dot{q}}{\partial p} - \frac{\partial L}{\partial \dot{q}}\frac{\partial \dot{q}}{\partial p} = \dot{q}^T, \\ \frac{\partial H}{\partial q} &= p^T \frac{\partial \dot{q}}{\partial q} - \frac{\partial L}{\partial q} - \frac{\partial L}{\partial \dot{q}}\frac{\partial \dot{q}}{\partial q} = -\frac{\partial L}{\partial q}. \end{aligned}$$

The Lagrange equations (1.4) are therefore equivalent to (1.7). □

Case of Quadratic *T*. In the case that $T = \frac{1}{2}\dot{q}^T M(q)\dot{q}$ is quadratic, where $M(q)$ is a symmetric and positive definite matrix, we have, for a fixed q, $p = M(q)\dot{q}$, so that the existence of the Legendre transform is established. Further, by replacing the variable $\dot{q}$ by $M(q)^{-1}p$ in the definition (1.6) of $H(p,q)$, we obtain

$$\begin{aligned} H(p,q) &= p^T M(q)^{-1}p - L\big(q, M(q)^{-1}p\big) \\ &= p^T M(q)^{-1}p - \frac{1}{2}p^T M(q)^{-1}p + U(q) = \frac{1}{2}p^T M(q)^{-1}p + U(q) \end{aligned}$$

and the Hamiltonian is $H = T + U$, which is the *total energy* of the mechanical system.

In Chap. I we have seen several examples of Hamiltonian systems, e.g., the pendulum (I.1.13), the Kepler problem (I.2.2), the outer solar system (I.2.12), etc. In the following we consider Hamiltonian systems (1.7) where the Hamiltonian $H(p,q)$ is arbitrary, and so not necessarily related to a mechanical problem.

VI.2 Symplectic Transformations

> The name "complex group" formerly advocated by me in allusion to line complexes, ... has become more and more embarrassing through collision with the word "complex" in the connotation of complex number. I therefore propose to replace it by the Greek adjective "symplectic."
>
> (H. Weyl (1939), p. 165)

A first property of Hamiltonian systems, already seen in Example 1.2 of Sect. IV.1, is that the Hamiltonian $H(p,q)$ is a *first integral* of the system (1.7). In this section we shall study another important property – the *symplecticity* of its flow. The basic objects to be studied are two-dimensional parallelograms lying in $\mathbb{R}^{2d}$. We suppose the parallelogram to be spanned by two vectors

$$\xi = \begin{pmatrix} \xi^p \\ \xi^q \end{pmatrix}, \qquad \eta = \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}$$

in the (p,q) space (ξ^p, ξ^q, η^p, η^q are in $\mathbb{R}^d$) as

$$P = \{t\xi + s\eta \mid 0 \le t \le 1,\ 0 \le s \le 1\}.$$

In the case $d = 1$ we consider the *oriented area*

$$\text{or.area}\,(P) = \det \begin{pmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{pmatrix} = \xi^p \eta^q - \xi^q \eta^p \tag{2.1}$$

(see left picture of Fig. 2.1). In higher dimensions, we replace this by the *sum of the oriented areas of the projections of P onto the coordinate planes (p_i, q_i),* i.e., by

$$\omega(\xi, \eta) := \sum_{i=1}^{d} \det \begin{pmatrix} \xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q \end{pmatrix} = \sum_{i=1}^{d} (\xi_i^p \eta_i^q - \xi_i^q \eta_i^p). \tag{2.2}$$

This defines a bilinear map acting on vectors of $\mathbb{R}^{2d}$, which will play a central role for Hamiltonian systems. In matrix notation, this map has the form

$$\omega(\xi, \eta) = \xi^T J \eta \qquad \text{with} \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{2.3}$$

where I is the identity matrix of dimension d.

Definition 2.1. A linear mapping $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called *symplectic* if

$$A^T J A = J$$

or, equivalently, if $\omega(A\xi, A\eta) = \omega(\xi, \eta)$ for all $\xi, \eta \in \mathbb{R}^{2d}$.

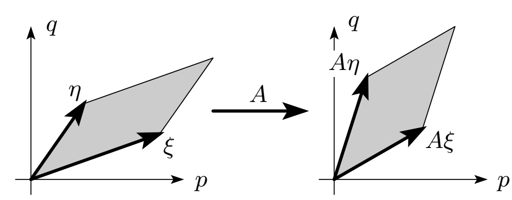


Fig. 2.1. Symplecticity (area preservation) of a linear mapping

In the case $d = 1$, where the expression $\omega(\xi, \eta)$ represents the area of the parallelogram P, symplecticity of a linear mapping A is therefore the *area preservation* of A (see Fig. 2.1). In the general case ($d > 1$), symplecticity means that the sum of the oriented areas of the projections of P onto (p_i, q_i) is the same as that for the transformed parallelograms $A(P)$.

We now turn our attention to nonlinear mappings. Differentiable functions can be locally approximated by linear mappings. This justifies the following definition.

Definition 2.2. A differentiable map $g : U \to \mathbb{R}^{2d}$ (where $U \subset \mathbb{R}^{2d}$ is an open set) is called *symplectic* if the Jacobian matrix $g'(p, q)$ is everywhere symplectic, i.e., if

$$g'(p, q)^T J\, g'(p, q) = J \qquad \text{or} \qquad \omega(g'(p, q)\xi, g'(p, q)\eta) = \omega(\xi, \eta).$$

Let us give a geometric interpretation of symplecticity for nonlinear mappings. Consider a 2-dimensional sub-manifold M of the $2d$-dimensional set U, and suppose that it is given as the image $M = \psi(K)$ of a compact set $K \subset \mathbb{R}^2$, where

$\psi(s,t)$ is a continuously differentiable function. The manifold M can then be considered as the limit of a union of small parallelograms spanned by the vectors

$$\frac{\partial\psi}{\partial s}(s,t)\,ds \qquad \text{and} \qquad \frac{\partial\psi}{\partial t}(s,t)\,dt.$$

For one such parallelogram we consider (as above) the sum over the oriented areas of its projections onto the (p_i, q_i) plane. We then sum over all parallelograms of the manifold. In the limit this gives the expression

$$\Omega(M) = \iint_K \omega\Big(\frac{\partial\psi}{\partial s}(s,t), \frac{\partial\psi}{\partial t}(s,t)\Big)\,ds\,dt. \tag{2.4}$$

The transformation formula for double integrals implies that $\Omega(M)$ is independent of the parametrization ψ of M.

Lemma 2.3. *If the mapping $g: U \to \mathbb{R}^{2d}$ is symplectic on U, then it preserves the expression $\Omega(M)$, i.e.,*

$$\Omega\big(g(M)\big) = \Omega(M)$$

holds for all 2-dimensional manifolds M that can be represented as the image of a continuously differentiable function ψ.

Proof. The manifold $g(M)$ can be parametrized by $g\circ\psi$. We have

$$\Omega\big(g(M)\big) = \iint_K \omega\Big(\frac{\partial(g\circ\psi)}{\partial s}(s,t), \frac{\partial(g\circ\psi)}{\partial t}(s,t)\Big)\,ds\,dt = \Omega(M),$$

because $(g\circ\psi)'(s,t) = g'\big(\psi(s,t)\big)\psi'(s,t)$ and g is a symplectic transformation. □

For $d = 1$, M is already a subset of $\mathbb{R}^2$ and we choose $K = M$ with ψ the identity map. In this case, $\Omega(M) = \iint_M ds\,dt$ represents the area of M. Hence, Lemma 2.3 states that all symplectic mappings (also nonlinear ones) are *area preserving*.

We are now able to prove the main result of this section. We use the notation $y = (p,q)$, and we write the Hamiltonian system (1.7) in the form

$$\dot{y} = J^{-1}\nabla H(y), \tag{2.5}$$

where J is the matrix of (2.3) and $\nabla H(y) = H'(y)^T$.

Recall that the flow $\varphi_t : U \to \mathbb{R}^{2d}$ of a Hamiltonian system is the mapping that advances the solution by time t, i.e., $\varphi_t(p_0,q_0) = (p(t,p_0,q_0), q(t,p_0,q_0))$, where $p(t,p_0,q_0)$, $q(t,p_0,q_0)$ is the solution of the system corresponding to initial values $p(0) = p_0$, $q(0) = q_0$.

Theorem 2.4 (Poincaré 1899). *Let $H(p,q)$ be a twice continuously differentiable function on $U \subset \mathbb{R}^{2d}$. Then, for each fixed t, the flow φ_t is a symplectic transformation wherever it is defined.*

Proof. The derivative $\partial\varphi_t/\partial y_0$ (with $y_0 = (p_0,q_0)$) is a solution of the variational equation which, for the Hamiltonian system (2.5), is of the form $\dot{\Psi} = J^{-1}\nabla^2 H\big(\varphi_t(y_0)\big)\Psi$, where $\nabla^2 H(p,q)$ is the Hessian matrix of $H(p,q)$ ($\nabla^2 H(p,q)$

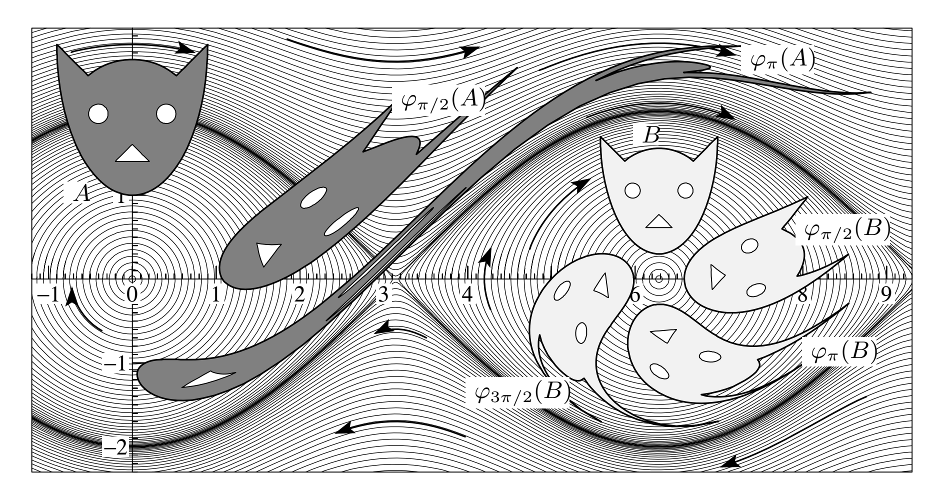


Fig. 2.2. Area preservation of the flow of Hamiltonian systems

is symmetric). We therefore obtain

$$\begin{aligned}\frac{d}{dt}\Big(\Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T J\Big(\frac{\partial\varphi_t}{\partial y_0}\Big)\Big) &= \Big(\frac{d}{dt}\frac{\partial\varphi_t}{\partial y_0}\Big)^T J\Big(\frac{\partial\varphi_t}{\partial y_0}\Big) + \Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T J\Big(\frac{d}{dt}\frac{\partial\varphi_t}{\partial y_0}\Big)\\ &= \Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T \nabla^2 H\big(\varphi_t(y_0)\big)J^{-T}J\Big(\frac{\partial\varphi_t}{\partial y_0}\Big) + \Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T \nabla^2 H\big(\varphi_t(y_0)\big)\Big(\frac{\partial\varphi_t}{\partial y_0}\Big) = 0,\end{aligned}$$

because $J^T = -J$ and $J^{-T}J = -I$. Since the relation

$$\Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T J\Big(\frac{\partial\varphi_t}{\partial y_0}\Big) = J \tag{2.6}$$

is satisfied for $t = 0$ (φ_0 is the identity map), it is satisfied for all t and all (p_0, q_0), as long as the solution remains in the domain of definition of H. □

Example 2.5. We illustrate this theorem with the pendulum problem (Example 1.2) using the normalization $m = \ell = g = 1$. We have $q = \alpha$, $p = \dot\alpha$, and the Hamiltonian is given by

$$H(p,q) = p^2/2 - \cos q.$$

Fig. 2.2 shows level curves of this function, and it also illustrates the area preservation of the flow φ_t. Indeed, by Theorem 2.4 and Lemma 2.3, the areas of A and $\varphi_t(A)$ as well as those of B and $\varphi_t(B)$ are the same, although their appearance is completely different.

We next show that symplecticity of the flow is a characteristic property for Hamiltonian systems. We call a differential equation $\dot y = f(y)$ *locally Hamiltonian*, if for every $y_0 \in U$ there exists a neighbourhood where $f(y) = J^{-1}\nabla H(y)$ for some function H.

Theorem 2.6. *Let $f : U \to \mathbb{R}^{2d}$ be continuously differentiable. Then, $\dot y = f(y)$ is locally Hamiltonian if and only if its flow $\varphi_t(y)$ is symplectic for all $y \in U$ and for all sufficiently small t.*

Proof. The necessity follows from Theorem 2.4. We therefore assume that the flow φ_t is symplectic, and we have to prove the local existence of a function $H(y)$ such that $f(y) = J^{-1}\nabla H(y)$. Differentiating (2.6) and using the fact that $\partial\varphi_t/\partial y_0$ is a solution of the variational equation $\dot{\Psi} = f'\big(\varphi_t(y_0)\big)\Psi$, we obtain

$$\frac{d}{dt}\bigg(\Big(\frac{\partial\varphi_t}{\partial y_0}\Big)^T J\Big(\frac{\partial\varphi_t}{\partial y_0}\Big)\bigg) = \Big(\frac{\partial\varphi_t}{\partial y_0}\Big)\Big(f'\big(\varphi_t(y_0)\big)^T J + Jf'\big(\varphi_t(y_0)\big)\Big)\Big(\frac{\partial\varphi_t}{\partial y_0}\Big) = 0.$$

Putting $t = 0$, it follows from $J = -J^T$ that $Jf'(y_0)$ is a symmetric matrix for all y_0. The Integrability Lemma 2.7 below shows that $Jf(y)$ can be written as the gradient of a function $H(y)$. □

The following integrability condition for the existence of a potential was already known to Euler and Lagrange (see e.g., Euler's *Opera Omnia*, vol. 19. p. 2-3, or Lagrange (1760), p. 375).

Lemma 2.7 (Integrability Lemma). *Let $D \subset \mathbb{R}^n$ be open and $f : D \to \mathbb{R}^n$ be continuously differentiable, and assume that the Jacobian $f'(y)$ is symmetric for all $y \in D$. Then, for every $y_0 \in D$ there exists a neighbourhood and a function $H(y)$ such that*

$$f(y) = \nabla H(y) \tag{2.7}$$

on this neighbourhood. In other words, the differential form $f_1(y)\,dy_1 + \ldots + f_n(y)\,dy_n = dH$ is a total differential.

Proof. Assume $y_0 = 0$, and consider a ball around y_0 which is contained in D. On this ball we define

$$H(y) = \int_0^1 y^T f(ty)\,dt + Const.$$

Differentiation with respect to y_k, and using the symmetry assumption $\partial f_i/\partial y_k = \partial f_k/\partial y_i$ yields

$$\frac{\partial H}{\partial y_k}(y) = \int_0^1 \Big(f_k(ty) + y^T\frac{\partial f}{\partial y_k}(ty)t\Big)\,dt = \int_0^1 \frac{d}{dt}\Big(tf_k(ty)\Big)\,dt = f_k(y),$$

which proves the statement. □

For $D = \mathbb{R}^{2d}$ or for star-shaped regions D, the above proof shows that the function H of Lemma 2.7 is globally defined. Hence the Hamiltonian of Theorem 2.6 is also globally defined in this case. This remains valid for simply connected sets D. A counter-example, which shows that the existence of a global Hamiltonian in Theorem 2.6 is not true for general D, is given in Exercise 6.

An important property of symplectic transformations, which goes back to Jacobi (1836, "Theorem X"), is that they preserve the Hamiltonian character of the differential equation. Such transformations have been termed *canonical* since the 19th century. The next theorem shows that canonical and symplectic transformations are the same.

Theorem 2.8. *Let $\psi : U \to V$ be a change of coordinates such that ψ and ψ^{-1} are continuously differentiable functions. If ψ is symplectic, the Hamiltonian system $\dot y = J^{-1}\nabla H(y)$ becomes in the new variables $z = \psi(y)$*

$$\dot z = J^{-1}\nabla K(z) \qquad \textit{with} \qquad K(z) = H(y). \tag{2.8}$$

Conversely, if ψ transforms every Hamiltonian system to another Hamiltonian system via (2.8), then ψ is symplectic.

Proof. Since $\dot z = \psi'(y)\dot y$ and $\psi'(y)^T\nabla K(z) = \nabla H(y)$, the Hamiltonian system $\dot y = J^{-1}\nabla H(y)$ becomes

$$\dot z = \psi'(y)J^{-1}\psi'(y)^T\nabla K(z) \tag{2.9}$$

in the new variables. It is equivalent to (2.8) if

$$\psi'(y)J^{-1}\psi'(y)^T = J^{-1}. \tag{2.10}$$

Multiplying this relation from the right by $\psi'(y)^{-T}$ and from the left by $\psi'(y)^{-1}$ and then taking its inverse yields $J = \psi'(y)^T J\psi'(y)$, which shows that (2.10) is equivalent to the symplecticity of ψ.

For the inverse relation we note that (2.9) is Hamiltonian for all $K(z)$ if and only if (2.10) holds. □

VI.3 First Examples of Symplectic Integrators

Since symplecticity is a characteristic property of Hamiltonian systems (Theorem 2.6), it is natural to search for numerical methods that share this property. Pioneering work on symplectic integration is due to de Vogelaere (1956), Ruth (1983), and Feng Kang (1985). Books on the now well-developed subject are Sanz-Serna & Calvo (1994) and Leimkuhler & Reich (2004).



Feng Kang[2]

Definition 3.1. A numerical one-step method is called *symplectic* if the one-step map

$$y_1 = \Phi_h(y_0)$$

is symplectic whenever the method is applied to a smooth Hamiltonian system.

[2] Feng Kang, born: 9 September 1920 in Nanjing (China), died: 17 August 1993 in Beijing; picture obtained from Yuming Shi with the help of Yifa Tang.

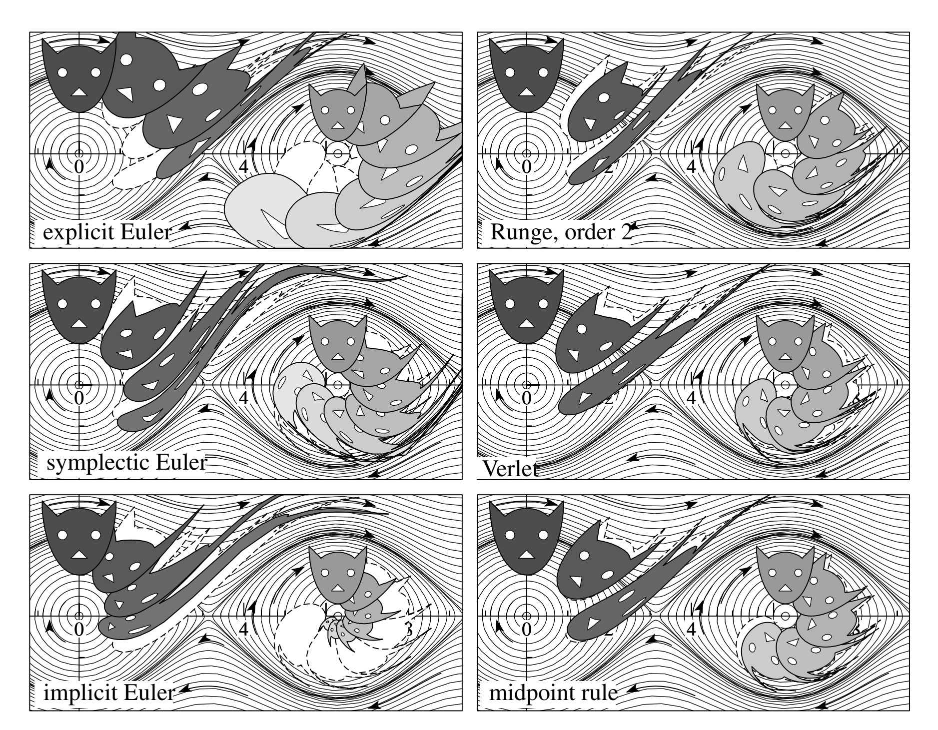


Fig. 3.1. Area preservation of numerical methods for the pendulum; same initial sets as in Fig. 2.2; first order methods (left column): $h = \pi/4$; second order methods (right column): $h = \pi/3$; dashed: exact flow

Example 3.2. We consider the pendulum problem of Example 2.5 with the same initial sets as in Fig. 2.2. We apply six different numerical methods to this problem: the explicit Euler method (I.1.5), the symplectic Euler method (I.1.9), and the implicit Euler method (I.1.6), as well as the second order method of Runge (II.1.3) (the right one), the Störmer–Verlet scheme (I.1.17), and the implicit midpoint rule (I.1.7). For two sets of initial values (p_0, q_0) we compute several steps with step size $h = \pi/4$ for the first order methods, and $h = \pi/3$ for the second order methods. One clearly observes in Fig. 3.1 that the explicit Euler, the implicit Euler and the second order explicit method of Runge are not symplectic (not area preserving). We shall prove below that the other methods are symplectic. A different proof of their symplecticity (using generating functions) will be given in Sect. VI.5.

In the following we show the symplecticity of various numerical methods from Chapters I and II when they are applied to the Hamiltonian system in the variables $y = (p, q)$,

$$\begin{aligned} \dot p &= -H_q(p,q) \\ \dot q &= H_p(p,q) \end{aligned} \qquad \text{or equivalently} \qquad \dot y = J^{-1}\nabla H(y),$$

where H_p and H_q denote the column vectors of partial derivatives of the Hamiltonian $H(p,q)$ with respect to p and q, respectively.

Theorem 3.3 (de Vogelaere 1956). *The so-called symplectic Euler methods (I.1.9)*

$$\begin{aligned} p_{n+1} &= p_n - hH_q(p_{n+1}, q_n) \\ q_{n+1} &= q_n + hH_p(p_{n+1}, q_n) \end{aligned} \qquad or \qquad \begin{aligned} p_{n+1} &= p_n - hH_q(p_n, q_{n+1}) \\ q_{n+1} &= q_n + hH_p(p_n, q_{n+1}) \end{aligned} \tag{3.1}$$

are symplectic methods of order 1.

Proof. We consider only the method to the left of (3.1). Differentiation with respect to (p_n, q_n) yields

$$\begin{pmatrix} I + hH_{qp}^T & 0 \\ -hH_{pp} & I \end{pmatrix} \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right) = \begin{pmatrix} I & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix},$$

where the matrices $H_{qp}, H_{pp}, \ldots$ of partial derivatives are all evaluated at (p_{n+1}, q_n). This relation allows us to compute $\frac{\partial(p_{n+1},q_{n+1})}{\partial(p_n,q_n)}$ and to check in a straightforward way the symplecticity condition $\big(\frac{\partial(p_{n+1},q_{n+1})}{\partial(p_n,q_n)}\big)^T J \big(\frac{\partial(p_{n+1},q_{n+1})}{\partial(p_n,q_n)}\big) = J$. □

The methods (3.1) are implicit for general Hamiltonian systems. For separable $H(p,q) = T(p) + U(q)$, however, both variants turn out to be explicit. It is interesting to mention that there are more general situations where the symplectic Euler methods are explicit. If, for a suitable ordering of the components,

$$\frac{\partial H}{\partial q_i}(p,q) \quad \text{does not depend on } p_j \text{ for } j \geq i, \tag{3.2}$$

then the left method of (3.1) is explicit, and the components of p_{n+1} can be computed one after the other. If, for a possibly different ordering of the components,

$$\frac{\partial H}{\partial p_i}(p,q) \quad \text{does not depend on } q_j \text{ for } j \geq i, \tag{3.3}$$

then the right method of (3.1) is explicit. As an example consider the Hamiltonian

$$H(p_r, p_\varphi, r, \varphi) = \frac{1}{2}\Big(p_r^2 + r^{-2}p_\varphi^2\Big) - r\cos\varphi + (r-1)^2,$$

which models a spring pendulum in polar coordinates. For the ordering $\varphi < r$, condition (3.2) is fulfilled, and for the inverse ordering $r < \varphi$ condition (3.3). Consequently, both symplectic Euler methods are explicit for this problem. The methods remain explicit if the conditions (3.2) and (3.3) hold for blocks of components instead of single components.

We consider next the extension of the Störmer–Verlet scheme (I.1.17), considered in Table II.2.1.

Theorem 3.4. *The Störmer–Verlet schemes (I.1.17)*

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2} H_q(p_{n+1/2}, q_n) \\ q_{n+1} &= q_n + \frac{h}{2}\Big(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})\Big) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2} H_q(p_{n+1/2}, q_{n+1}) \end{aligned} \tag{3.4}$$

and

$$
\begin{aligned}
q_{n+1/2} &= q_n + \frac{h}{2} H_q(p_n, q_{n+1/2}) \\
p_{n+1} &= p_n - \frac{h}{2}\Big(H_p(p_n, q_{n+1/2}) + H_p(p_{n+1}, q_{n+1/2}) \Big) \\
q_{n+1} &= q_{n+1/2} + \frac{h}{2} H_q(p_{n+1}, q_{n+1/2})
\end{aligned}
\tag{3.5}
$$

are symplectic methods of order 2.

Proof. This is an immediate consequence of the fact that the Störmer–Verlet scheme is the composition of the two symplectic Euler methods (3.1). Order 2 follows from its symmetry. □

We note that the Störmer–Verlet methods (3.4) and (3.5) are explicit for separable problems and for Hamiltonians that satisfy both conditions (3.2) and (3.3).

Theorem 3.5. *The implicit midpoint rule*

$$
y_{n+1} = y_n + hJ^{-1}\nabla H\big((y_{n+1} + y_n)/2\big) \tag{3.6}
$$

is a symplectic method of order 2.

Proof. Differentiation of (3.6) yields

$$
\Big(I - \frac{h}{2} J^{-1}\nabla^2 H\Big)\Big(\frac{\partial y_{n+1}}{\partial y_n}\Big) = \Big(I + \frac{h}{2} J^{-1}\nabla^2 H\Big).
$$

Again it is straightforward to verify that $\big(\frac{\partial y_{n+1}}{\partial y_n}\big)^T J \big(\frac{\partial y_{n+1}}{\partial y_n}\big) = J$. Due to its symmetry, the midpoint rule is known to be of order 2 (see Theorem II.3.2). □

The next two theorems are a consequence of the fact that the composition of symplectic transformations is again symplectic. They are also used to prove the existence of symplectic methods of arbitrarily high order, and to explain why the theory of composition methods of Chapters II and III is so important for geometric integration.

Theorem 3.6. *Let Φ_h denote the symplectic Euler method (3.1). Then, the composition method (II.4.6) is symplectic for every choice of the parameters α_i, β_i.*

If $\widehat{\Phi}_h$ is symplectic and symmetric (e.g., the implicit midpoint rule or the Störmer–Verlet scheme), then the composition method (V.3.8) is symplectic too. □

Theorem 3.7. *Assume that the Hamiltonian is given by $H(y) = H_1(y) + H_2(y)$, and consider the splitting*

$$
\dot y = J^{-1}\nabla H(y) = J^{-1}\nabla H_1(y) + J^{-1}\nabla H_2(y).
$$

The splitting method (II.5.6) is then symplectic. □

VI.4 Symplectic Runge–Kutta Methods

The systematic study of symplectic Runge–Kutta methods started around 1988, and a complete characterization has been found independently by Lasagni (1988) (using the approach of generating functions), and by Sanz-Serna (1988) and Suris (1988) (using the ideas of the classical papers of Burrage & Butcher (1979) and Crouzeix (1979) on algebraic stability).

VI.4.1 Criterion of Symplecticity

We follow the approach of Bochev & Scovel (1994), which is based on the following important lemma.

Lemma 4.1. *For Runge–Kutta methods and for partitioned Runge–Kutta methods the following diagram commutes:*

$$\begin{array}{ccc} \dot y = f(y),\;\; y(0)=y_0 & \longrightarrow & \begin{array}{c}\dot y = f(y),\;\; y(0)=y_0\\ \dot\Psi = f'(y)\Psi,\;\; \Psi(0)=I\end{array} \\ \Big\downarrow \textit{method} & & \Big\downarrow \textit{method} \\ \{y_n\} & \longrightarrow & \{y_n,\Psi_n\} \end{array}$$

(horizontal arrows mean a differentiation with respect to y_0). Therefore, the numerical result y_n, Ψ_n, obtained from applying the method to the problem augmented by its variational equation, is equal to the numerical solution for $\dot y = f(y)$ augmented by its derivative $\Psi_n = \partial y_n/\partial y_0$.

Proof. The result is proved by implicit differentiation. Let us illustrate this for the explicit Euler method

$$y_{n+1} = y_n + hf(y_n).$$

We consider y_n and y_{n+1} as functions of y_0, and we differentiate with respect to y_0 the equation defining the numerical method. For the Euler method this gives

$$\frac{\partial y_{n+1}}{\partial y_0} = \frac{\partial y_n}{\partial y_0} + hf'(y_n)\,\frac{\partial y_n}{\partial y_0},$$

which is exactly the relation that we get from applying the method to the variational equation. Since $\partial y_0/\partial y_0 = I$, we have $\partial y_n/\partial y_0 = \Psi_n$ for all n. □

The main observation now is that the symplecticity condition (2.6) is a quadratic first integral of the variational equation: we write the Hamiltonian system together with its variational equation as

$$\dot y = J^{-1}\nabla H(y), \qquad \dot\Psi = J^{-1}\nabla^2 H(y)\Psi. \tag{4.1}$$

It follows from

$$(J^{-1}\nabla^2 H(y)\Psi)^T J\Psi + \Psi^T J(J^{-1}\nabla^2 H(y)\Psi) = 0$$

(see also the proof of Theorem 2.4) that $\Psi^T J\Psi$ is a quadratic first integral of the augmented system (4.1).

Therefore, every Runge–Kutta method that preserves quadratic first integrals, is a symplectic method. From Theorem IV.2.1 and Theorem IV.2.2 we thus obtain the following results.

Theorem 4.2. *The Gauss collocation methods of Sect. II.1.3 are symplectic.* □

Theorem 4.3. *If the coefficients of a Runge–Kutta method satisfy*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \qquad \textit{for all } \; i,j = 1,\ldots,s, \tag{4.2}$$

then it is symplectic. □

Similar to the situation in Theorem V.2.4, diagonally implicit, symplectic Runge–Kutta methods are composition methods.

Theorem 4.4. *A diagonally implicit Runge–Kutta method satisfying the symplecticity condition (4.2) and $b_i \neq 0$ is equivalent to the composition*

$$\Phi^M_{b_s h} \circ \ldots \circ \Phi^M_{b_2 h} \circ \Phi^M_{b_1 h},$$

where Φ^M_h stands for the implicit midpoint rule.

Proof. For $i = j$ condition (4.2) gives $a_{ii} = b_i/2$ and, together with $a_{ji} = 0$ (for $i > j$), implies $a_{ij} = b_j$. This proves the statement. □

The assumption "$b_i \neq 0$" is not restrictive in the sense that for diagonally implicit Runge–Kutta methods satisfying (4.2) the internal stages corresponding to "$b_i = 0$" do not influence the numerical result and can be removed.

To understand the symplecticity of partitioned Runge–Kutta methods, we write the solution Ψ of the variational equation as

$$\Psi = \begin{pmatrix} \Psi^p \\ \Psi^q \end{pmatrix}.$$

Then, the Hamiltonian system together with its variational equation (4.1) is a partitioned system with variables (p, Ψ^p) and (q, Ψ^q). Every component of

$$\Psi^T J\Psi = (\Psi^p)^T \Psi^q - (\Psi^q)^T \Psi^p$$

is of the form (IV.2.5), so that Theorem IV.2.3 and Theorem IV.2.4 yield the following results.

Theorem 4.5. *The Lobatto IIIA - IIIB pair is a symplectic method.* □

Theorem 4.6. *If the coefficients of a partitioned Runge–Kutta method (II.2.2) satisfy*

$$b_i \widehat{a}_{ij} + \widehat{b}_j a_{ji} = b_i \widehat{b}_j \qquad \text{for } i, j = 1, \ldots, s, \tag{4.3}$$

$$b_i = \widehat{b}_i \qquad \text{for } i = 1, \ldots, s, \tag{4.4}$$

then it is symplectic.

If the Hamiltonian is of the form $H(p,q) = T(p) + U(q)$, i.e., it is separable, then the condition (4.3) alone implies the symplecticity of the numerical flow. □

We have seen in Sect. V.2.2 that within the class of partitioned Runge–Kutta methods it is possible to get explicit, symmetric methods for separable systems $\dot{y} = f(z)$, $\dot{z} = g(y)$. A similar result holds for symplectic methods. However, as in Theorem V.2.6, such methods are not more general than composition or splitting methods as considered in Sect. II.5. This has first been observed by Okunbor & Skeel (1992).

Theorem 4.7. *Consider a partitioned Runge–Kutta method based on two diagonally implicit methods (i.e., $a_{ji} = \widehat{a}_{ji} = 0$ for $i > j$), assume $a_{ii} \cdot \widehat{a}_{ii} = 0$ for all i, and apply it to a separable Hamiltonian system with $H(p,q) = T(p) + U(q)$. If (4.3) holds, then the numerical result is the same as that obtained from the splitting method (II.5.6).*

By (II.5.8), such a method is equivalent to a composition of symplectic Euler steps.

Proof. We first notice that the stage values $k_i = f(Z_i)$ (for i with $b_i = 0$) and $\ell_i = g(Y_i)$ (for i with $\widehat{b}_i = 0$) do not influence the numerical solution and can be removed. This yields a scheme with non-zero b_i and $\widehat{b}_i$, but with possibly non-square matrices (a_{ij}) and $(\widehat{a}_{ij})$.

Since the method is explicit for separable problems, one of the reduced matrices (a_{ij}) or $(\widehat{a}_{ij})$ has a row consisting only of zeros. Assume that it is the first row of (a_{ij}), so that $a_{1j} = 0$ for all j. The symplecticity condition thus implies $\widehat{a}_{i1} = \widehat{b}_1 \neq 0$ for all $i \geq 1$, and $a_{i1} = b_1 \neq 0$ for $i \geq 2$. This then yields $\widehat{a}_{22} \neq 0$, because otherwise the first two stages of $(\widehat{a}_{ij})$ would be identical and one could be removed. By our assumption we get $a_{22} = 0$, $\widehat{a}_{i2} = \widehat{b}_2 \neq 0$ for $i \geq 2$, and $a_{i2} = b_2$ for $i \geq 3$. Continuing this procedure we see that the method becomes

$$\ldots \circ \varphi^{[2]}_{\widehat{b}_2 h} \circ \varphi^{[1]}_{b_2 h} \circ \varphi^{[2]}_{\widehat{b}_1 h} \circ \varphi^{[1]}_{b_1 h},$$

where $\varphi^{[1]}_t$ and $\varphi^{[2]}_t$ are the exact flows corresponding to the Hamiltonians $T(p)$ and $U(q)$, respectively. □

The necessity of the conditions of Theorem 4.3 and Theorem 4.6 for symplectic (partitioned) Runge–Kutta methods will be discussed at the end of this chapter in Sect. VI.7.4.

A second order differential equation $\ddot{y} = g(y)$, augmented by its variational equation, is again of this special form. Furthermore, the diagram of Lemma 4.1 commutes for Nyström methods, so that Theorem IV.2.5 yields the following result originally obtained by Suris (1988, 1989).

Theorem 4.8. *If the coefficients of a Nyström method (IV.2.11) satisfy*

$$\begin{aligned} \beta_i &= b_i(1-c_i) \qquad \textit{for } i=1,\ldots,s, \\ b_i(\beta_j - a_{ij}) &= b_j(\beta_i - a_{ji}) \qquad \textit{for } i,j=1,\ldots,s, \end{aligned} \tag{4.5}$$

then it is symplectic. □

VI.4.2 Connection Between Symplectic and Symmetric Methods

There exist symmetric methods that are not symplectic, and there exist symplectic methods that are not symmetric. For example, the *trapezoidal rule*

$$y_1 = y_0 + \frac{h}{2}\Big(f(y_0) + f(y_1)\Big) \tag{4.6}$$

is symmetric, but it does not satisfy the condition (4.2) for symplecticity. In fact, this is true of all Lobatto IIIA methods (see Example II.2.2). On the other hand, any composition $\Phi_{\gamma_1 h} \circ \Phi_{\gamma_2 h}$ ($\gamma_1 + \gamma_2 = 1$) of symplectic methods is symplectic but symmetric only if $\gamma_1 = \gamma_2$.

However, for (non-partitioned) Runge–Kutta methods and for quadratic Hamiltonians $H(y) = \frac{1}{2} y^T C y$ (C is a symmetric real matrix), where the corresponding system (2.5) is linear,

$$\dot{y} = J^{-1} C y, \tag{4.7}$$

we shall see that both concepts are equivalent.

A Runge–Kutta method, applied with step size h to a linear system $\dot{y} = Ly$, is equivalent to

$$y_1 = R(hL) y_0, \tag{4.8}$$

where the rational function $R(z)$ is given by

$$R(z) = 1 + z b^T (I - zA)^{-1} \mathbb{1}, \tag{4.9}$$

$A = (a_{ij})$, $b^T = (b_1, \ldots, b_s)$, and $\mathbb{1}^T = (1, \ldots, 1)$. The function $R(z)$ is called the *stability function* of the method, and it is familiar to us from the study of stiff differential equations (see e.g., Hairer & Wanner (1996), Chap. IV.3).

For the explicit Euler method, the implicit Euler method and the implicit midpoint rule, the stability function $R(z)$ is given by

$$1 + z, \qquad \frac{1}{1-z}, \qquad \frac{1 + z/2}{1 - z/2}.$$

Theorem 4.9. *For Runge–Kutta methods the following statements are equivalent:*
- *the method is symmetric for linear problems $\dot{y} = Ly$;*
- *the method is symplectic for problems (4.7) with symmetric C;*
- *the stability function satisfies $R(-z)R(z) = 1$ for all complex z.*

Proof. The method $y_1 = R(hL)y_0$ is symmetric, if and only if $y_0 = R(-hL)y_1$ holds for all initial values y_0. But this is equivalent to $R(-hL)R(hL) = I$.

Since $\Phi_h'(y_0) = R(hL)$, symplecticity of the method for the problem (4.7) is defined by $R(hJ^{-1}C)^T JR(hJ^{-1}C) = J$. For $R(z) = P(z)/Q(z)$ this is equivalent to

$$P(hJ^{-1}C)^T JP(hJ^{-1}C) = Q(hJ^{-1}C)^T JQ(hJ^{-1}C). \tag{4.10}$$

By the symmetry of C, the matrix $L := J^{-1}C$ satisfies $L^T J = -JL$ and hence also $(L^k)^T J = J(-L)^k$ for $k = 0, 1, 2, \ldots$. Consequently, (4.10) is equivalent to

$$P(-hJ^{-1}C)P(hJ^{-1}C) = Q(-hJ^{-1}C)Q(hJ^{-1}C),$$

which is nothing other than $R(-hJ^{-1}C)R(hJ^{-1}C) = I$. □

VI.5 Generating Functions

> ... by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function. (W.R. Hamilton 1834)

> Professor Hamilton hat ... das merkwürdige Resultat gefunden, dass ... sich die Integralgleichungen der Bewegung ... sämmtlich durch die partiellen Differentialquotienten einer einzigen Function darstellen lassen. (C.G.J. Jacobi 1837)

We enter here the second heaven of Hamiltonian theory, the realm of partial differential equations and generating functions. The starting point of this theory was the discovery of Hamilton that the motion of the system is completely described by a "characteristic" function S, and that S is the solution of a partial differential equation, now called the *Hamilton–Jacobi differential equation*.

It was noticed later, especially by Siegel (see Siegel & Moser 1971, §3), that such a function S is directly connected to any symplectic map. It received the name *generating function*.

VI.5.1 Existence of Generating Functions

We now consider a fixed Hamiltonian system and a fixed time interval and denote by the column vectors p and q the *initial values* $p_1, \ldots, p_d$ and $q_1, \ldots, q_d$ at t_0 of a trajectory. The *final values* at t_1 are written as P and Q. We thus have a mapping $(p, q) \mapsto (P, Q)$ which, as we know, is symplectic on an open set U.

The following results are conveniently formulated in the notation of differential forms. For a function F we denote by $dF = F'$ its (Fréchet) derivative. We denote by $dq = (dq_1, \ldots, dq_d)^T$ the derivative of the coordinate projection $(p, q) \mapsto q$.

Theorem 5.1. *A mapping $\varphi : (p,q) \mapsto (P,Q)$ is symplectic if and only if there exists locally a function $S(p,q)$ such that*

$$P^T\,dQ - p^T\,dq = dS. \tag{5.1}$$

This means that $P^T\,dQ - p^T\,dq$ is a total differential.

Proof. We split the Jacobian of φ into the natural 2×2 block matrix

$$\frac{\partial(P,Q)}{\partial(p,q)} = \begin{pmatrix} P_p & P_q \\ Q_p & Q_q \end{pmatrix}.$$

Inserting this into (2.6) and multiplying out shows that the three conditions

$$P_p^T Q_p = Q_p^T P_p, \qquad P_p^T Q_q - I = Q_p^T P_q, \qquad Q_q^T P_q = P_q^T Q_q \tag{5.2}$$

are equivalent to symplecticity. We now insert $dQ = Q_p\,dp + Q_q\,dq$ into the left-hand side of (5.1) and obtain

$$\Big(P^T Q_p,\, P^T Q_q - p^T\Big)\begin{pmatrix} dp \\ dq \end{pmatrix} = \begin{pmatrix} Q_p^T P \\ Q_q^T P - p \end{pmatrix}^T \begin{pmatrix} dp \\ dq \end{pmatrix}.$$

To apply the Integrability Lemma 2.7, we just have to verify the symmetry of the Jacobian of the coefficient vector,

$$\begin{pmatrix} Q_p^T P_p & Q_p^T P_q \\ Q_q^T P_p - I & Q_q^T P_q \end{pmatrix} + \sum_i P_i\,\frac{\partial^2 Q_i}{\partial(p,q)^2}. \tag{5.3}$$

Since the Hessians of Q_i are symmetric anyway, it is immediately clear that the symmetry of the matrix (5.3) is equivalent to the symplecticity conditions (5.2). □

Reconstruction of the Symplectic Map from S. Up to now we have considered all functions as depending on p and q. The essential idea now is to introduce new coordinates; namely (5.1) suggests using $z = (q,Q)$ instead of $y = (p,q)$. This is a well-defined local change of coordinates $y = \psi(z)$ if p can be expressed in terms of the coordinates (q,Q), which is possible by the implicit function theorem if $\frac{\partial Q}{\partial p}$ is invertible. Abusing our notation we again write $S(q,Q)$ for the transformed function $S(\psi(z))$. Then, by comparing the coefficients of $dS = \frac{\partial S(q,Q)}{\partial q}\,dq + \frac{\partial S(q,Q)}{\partial Q}\,dQ$ with (5.1), we arrive at[3]

$$P = \frac{\partial S}{\partial Q}(q,Q), \qquad p = -\frac{\partial S}{\partial q}(q,Q). \tag{5.4}$$

If the transformation $(p,q) \mapsto (P,Q)$ is symplectic, then it can be reconstructed from the scalar function $S(q,Q)$ by the relations (5.4). By Theorem 5.1 the converse

[3] On the right-hand side we should have put the gradient $\nabla_Q S = (\partial S/\partial Q)^T$. We shall not make this distinction between row and column vectors when there is no danger of confusion.

is also true: *any sufficiently smooth and nondegenerate function $S(q,Q)$ "generates" via (5.4) a symplectic mapping $(p,q) \mapsto (P,Q)$*. This gives us a powerful tool for creating symplectic methods.

Mixed-Variable Generating Functions. Another often useful choice of coordinates for generating symplectic maps are the mixed variables (P,q). For any continuously differentiable function $\widehat{S}(P,q)$ we clearly have $d\widehat{S} = \frac{\partial \widehat{S}}{\partial P}\, dP + \frac{\partial \widehat{S}}{\partial q}\, dq$. On the other hand, since $d(P^T Q) = P^T dQ + Q^T dP$, the symplecticity condition (5.1) can be rewritten as $Q^T dP + p^T dq = d(Q^T P - S)$ for some function S. It therefore follows from Theorem 5.1 that the equations

$$Q = \frac{\partial \widehat{S}}{\partial P}(P,q), \qquad p = \frac{\partial \widehat{S}}{\partial q}(P,q) \tag{5.5}$$

define (locally) a symplectic map $(p,q) \mapsto (P,Q)$ if $\partial^2 \widehat{S}/\partial P \partial q$ is invertible.

Example 5.2. Let $Q = \chi(q)$ be a change of position coordinates. With the generating function $\widehat{S}(P,q) = P^T \chi(q)$ we obtain via (5.5) an extension to a symplectic mapping $(p,q) \mapsto (P,Q)$. The conjugate variables are thus related by $p = \chi'(q)^T P$.

Mappings Close to the Identity. We are mainly interested in the situation where the mapping $(p,q) \mapsto (P,Q)$ is close to the identity. In this case, the choices (p,Q) or (P,q) or $\big((P+p)/2, (Q+q)/2\big)$ of independent variables are convenient and lead to the following characterizations.

Lemma 5.3. *Let $(p,q) \mapsto (P,Q)$ be a smooth transformation, close to the identity. It is symplectic if and only if one of the following conditions holds locally:*

- $Q^T dP + p^T dq = d(P^T q + S^1)$ *for some function* $S^1(P,q)$*;*
- $P^T dQ + q^T dp = d(p^T Q - S^2)$ *for some function* $S^2(p,Q)$*;*
- $(Q-q)^T d(P+p) - (P-p)^T d(Q+q) = 2\, dS^3$
  for some function $S^3\big((P+p)/2, (Q+q)/2\big)$.

Proof. The first characterization follows from the discussion before formula (5.5) if we put S^1 such that $P^T q + S^1 = \widehat{S} = Q^T P - S$. For the second characterization we use $d(p^T q) = p^T dq + q^T dp$ and the same arguments as before. The last one follows from the fact that (5.1) is equivalent to $(Q-q)^T d(P+p) - (P-p)^T d(Q+q) = d\big((P+p)^T (Q-q) - 2S\big)$. □

The generating functions S^1, S^2, and S^3 have been chosen such that we obtain the identity mapping when they are replaced with zero. Comparing the coefficient functions of dq and dP in the first characterization of Lemma 5.3, we obtain

$$p = P + \frac{\partial S^1}{\partial q}(P,q), \qquad Q = q + \frac{\partial S^1}{\partial P}(P,q). \tag{5.6}$$

Whatever the scalar function $S^1(P,q)$ is, the relation (5.6) defines a symplectic transformation $(p,q) \mapsto (P,Q)$. For $S^1(P,q) := hH(P,q)$ we recognize the symplectic Euler method (I.1.9). This is an elegant proof of the symplecticity of this method. The second characterization leads to the adjoint of the symplectic Euler method.

The third characterization of Lemma 5.3 can be written as

$$\begin{aligned} P &= p - \partial_2 S^3\big((P+p)/2, (Q+q)/2\big), \\ Q &= q + \partial_1 S^3\big((P+p)/2, (Q+q)/2\big), \end{aligned} \tag{5.7}$$

which, for $S^3 = hH$, is nothing other than the implicit midpoint rule (I.1.7) applied to a Hamiltonian system. We have used the notation ∂_1 and ∂_2 for the derivative with respect to the first and second argument, respectively. The system (5.7) can also be written in compact form as

$$Y = y + J^{-1}\nabla S^3\big((Y+y)/2\big), \tag{5.8}$$

where $Y = (P,Q)$, $y = (p,q)$, $S^3(w) = S^3(u,v)$ with $w = (u,v)$, and J is the matrix of (2.3).

VI.5.2 Generating Function for Symplectic Runge–Kutta Methods

We have just seen that all symplectic transformations can be written in terms of generating functions. What are these generating functions for symplectic Runge–Kutta methods? The following result, proved by Lasagni in an unpublished manuscript (with the same title as the note Lasagni (1988)), gives an alternative proof for Theorem 4.3.

Theorem 5.4. *Suppose that*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \textit{for all } \; i,j \tag{5.9}$$

(see Theorem 4.3). Then, the Runge–Kutta method

$$\begin{aligned} P &= p - h\sum_{i=1}^s b_i H_q(P_i,Q_i), \qquad & P_i &= p - h\sum_{j=1}^s a_{ij} H_q(P_j,Q_j), \\ Q &= q + h\sum_{i=1}^s b_i H_p(P_i,Q_i), \qquad & Q_i &= q + h\sum_{j=1}^s a_{ij} H_p(P_j,Q_j) \end{aligned} \tag{5.10}$$

can be written as (5.6) with

$$S^1(P,q,h) = h\sum_{i=1}^s b_i H(P_i,Q_i) - h^2 \sum_{i,j=1}^s b_i a_{ij} H_q(P_i,Q_i)^T H_p(P_j,Q_j). \tag{5.11}$$

Proof. We first differentiate $S^1(P,q,h)$ with respect to q. Using the abbreviations $H[i]=H(P_i,Q_i)$, $H_p[i]=H_p(P_i,Q_i),\ldots$, we obtain

$$\begin{aligned}\frac{\partial}{\partial q}\Big(\sum_i b_i H[i]\Big) &= \sum_i b_i H_p[i]^T\Big(\frac{\partial p}{\partial q}-h\sum_j a_{ij}\frac{\partial}{\partial q}H_q[j]\Big)\\ &+ \sum_i b_i H_q[i]^T\Big(I+h\sum_j a_{ij}\frac{\partial}{\partial q}H_p[j]\Big).\end{aligned}$$

With

$$0=\frac{\partial p}{\partial q}-h\sum_j b_j\frac{\partial}{\partial q}H_q[j]$$

(this is obtained by differentiating the first relation of (5.10)), Leibniz' rule

$$\frac{\partial}{\partial q}\big(H_q[i]^T H_p[j]\big)=H_q[i]^T\frac{\partial}{\partial q}H_p[j]+H_p[j]^T\frac{\partial}{\partial q}H_q[i]$$

and the condition (5.9) therefore yield the first relation of

$$\frac{\partial S^1(P,q,h)}{\partial q}=h\sum_i b_i H_q[i],\qquad \frac{\partial S^1(P,q,h)}{\partial P}=h\sum_i b_i H_p[i].$$

The second relation is proved in the same way. This shows that the Runge–Kutta formulas (5.10) are equivalent to (5.6). □

It is interesting to note that, whereas Lemma 5.3 guarantees the *local* existence of a generating function S^1, the explicit formula (5.11) shows that for Runge–Kutta methods this generating function is *globally* defined. This means that it is well-defined in the same region where the Hamiltonian $H(p,q)$ is defined.

Theorem 5.5. *A partitioned Runge–Kutta method (II.2.2), satisfying the symplecticity conditions (4.3) and (4.4), is equivalent to (5.6) with*

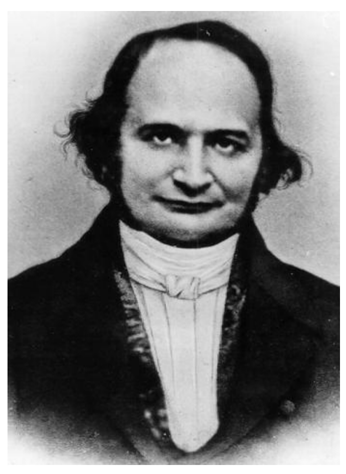
$$S^1(P,q,h)=h\sum_{i=1}^s b_i H(P_i,Q_i)-h^2\sum_{i,j=1}^s b_i\widehat{a}_{ij}H_q(P_i,Q_i)^T H_p(P_j,Q_j).$$

If the Hamiltonian is of the form $H(p,q)=T(p)+U(q)$, i.e., it is separable, then the condition (4.3) alone implies that the method is of the form (5.6) with

$$S^1(P,q,h)=h\sum_{i=1}^s\Big(b_iU(Q_i)+\widehat{b}_iT(P_i)\Big)-h^2\sum_{i,j=1}^s b_i\widehat{a}_{ij}U_q(Q_i)^T T_p(P_j,).$$

Proof. This is a straightforward extension of the proof of the previous theorem. □

VI.5.3 The Hamilton–Jacobi Partial Differential Equation



C.G.J. Jacobi[4]

We now return to the above construction of S for a symplectic transformation $(p,q) \mapsto (P,Q)$ (see Theorem 5.1). This time, however, we imagine the point $P(t), Q(t)$ to move in the flow of the Hamiltonian system (1.7). We wish to determine a smooth generating function $S(q,Q,t)$, now also depending on t, which generates via (5.4) the symplectic map $(p,q) \mapsto \big(P(t),Q(t)\big)$ of the *exact flow* of the Hamiltonian system.

In accordance with equation (5.4) we have to satisfy

$$\begin{aligned} P_i(t) &= \frac{\partial S}{\partial Q_i}\big(q,Q(t),t\big), \\ p_i &= -\frac{\partial S}{\partial q_i}\big(q,Q(t),t\big). \end{aligned} \tag{5.12}$$

Differentiating the second relation with respect to t yields

$$0 = \frac{\partial^2 S}{\partial q_i \partial t}\big(q,Q(t),t\big) + \sum_{j=1}^{d} \frac{\partial^2 S}{\partial q_i \partial Q_j}\big(q,Q(t),t\big)\cdot \dot Q_j(t) \tag{5.13}$$

$$= \frac{\partial^2 S}{\partial q_i \partial t}\big(q,Q(t),t\big) + \sum_{j=1}^{d} \frac{\partial^2 S}{\partial q_i \partial Q_j}\big(q,Q(t),t\big)\cdot \frac{\partial H}{\partial P_j}\big(P(t),Q(t)\big) \tag{5.14}$$

where we have inserted the second equation of (1.7) for $\dot Q_j$. Then, using the chain rule, this equation simplifies to

$$\frac{\partial}{\partial q_i}\left(\frac{\partial S}{\partial t} + H\Big(\frac{\partial S}{\partial Q_1},\ldots,\frac{\partial S}{\partial Q_d}, Q_1,\ldots,Q_d\Big)\right) = 0. \tag{5.15}$$

This motivates the following surprisingly simple relation.

Theorem 5.6. *If $S(q,Q,t)$ is a smooth solution of the partial differential equation*

$$\frac{\partial S}{\partial t} + H\Big(\frac{\partial S}{\partial Q_1},\ldots,\frac{\partial S}{\partial Q_d}, Q_1,\ldots,Q_d\Big) = 0 \tag{5.16}$$

with initial values satisfying $\frac{\partial S}{\partial q_i}(q,q,0) + \frac{\partial S}{\partial Q_i}(q,q,0) = 0$, and if the matrix $\big(\frac{\partial^2 S}{\partial q_i \partial Q_j}\big)$ is invertible, then the map $(p,q) \mapsto \big(P(t),Q(t)\big)$ defined by (5.12) is the flow $\varphi_t(p,q)$ of the Hamiltonian system (1.7).

Equation (5.16) is called the "Hamilton–Jacobi partial differential equation".

[4] Carl Gustav Jacob Jacobi, born: 10 December 1804 in Potsdam (near Berlin), died: 18 February 1851 in Berlin.

Proof. The invertibility of the matrix $\big(\frac{\partial^2 S}{\partial q_i \partial Q_j}\big)$ and the implicit function theorem imply that the mapping $(p,q) \mapsto \big(P(t),Q(t)\big)$ is well-defined by (5.12), and, by differentiation, that (5.13) is true as well.

Since, by hypothesis, $S(q,Q,t)$ is a solution of (5.16), the equations (5.15) and hence also (5.14) are satisfied. Subtracting (5.13) and (5.14), and once again using the invertibility of the matrix $\big(\frac{\partial^2 S}{\partial q_i \partial Q_j}\big)$, we see that necessarily $\dot{Q}(t) = H_p\big(P(t),Q(t)\big)$. This proves the validity of the second equation of the Hamiltonian system (1.7).

The first equation of (1.7) is obtained as follows: differentiate the first relation of (5.12) with respect to t and the Hamilton–Jacobi equation (5.16) with respect to Q_i, then eliminate the term $\frac{\partial^2 S}{\partial Q_i \partial t}$. Using $\dot{Q}(t) = H_p\big(P(t),Q(t)\big)$, this leads in a straightforward way to $\dot{P}(t) = -H_q\big(P(t),Q(t)\big)$. The condition on the initial values of S ensures that $(P(0),Q(0)) = (p,q)$. □

In the hands of Jacobi (1842), this equation turned into a powerful tool for the analytic integration of many difficult problems. One has, in fact, to find a solution of (5.16) which contains sufficiently many parameters. This is often possible with the method of separation of variables. An example is presented in Exercise 11.

Hamilton–Jacobi Equation for S^1, S^2, and S^3. We now express the Hamilton–Jacobi differential equation in the coordinates used in Lemma 5.3. In these coordinates it is also possible to prescribe initial values for S at $t=0$.

From the proof of Lemma 5.3 we know that the generating functions in the variables (q,Q) and (P,q) are related by

$$S^1(P,q,t) = P^T(Q-q) - S(q,Q,t). \tag{5.17}$$

We consider P,q,t as independent variables, and we differentiate this relation with respect to t. Using the first relation of (5.12) this gives

$$\frac{\partial S^1}{\partial t}(P,q,t) = P^T \frac{\partial Q}{\partial t} - \frac{\partial S}{\partial Q}(q,Q,t)\frac{\partial Q}{\partial t} - \frac{\partial S}{\partial t}(q,Q,t) = -\frac{\partial S}{\partial t}(q,Q,t).$$

Differentiating (5.17) with respect to P yields

$$\frac{\partial S^1}{\partial P}(P,q,t) = Q - q + P^T\frac{\partial Q}{\partial P} - \frac{\partial S}{\partial Q}(q,Q,t)\frac{\partial Q}{\partial P} = Q - q.$$

Inserting $\frac{\partial S}{\partial Q} = P$ and $Q = q + \frac{\partial S^1}{\partial P}$ into the Hamilton–Jacobi equation (5.16) we are led to the equation of the following theorem.

Theorem 5.7. *If $S^1(P,q,t)$ is a solution of the partial differential equation*

$$\frac{\partial S^1}{\partial t}(P,q,t) \;=\; H\Big(P, q + \frac{\partial S^1}{\partial P}(P,q,t)\Big), \qquad S^1(P,q,0) = 0, \tag{5.18}$$

then the mapping $(p,q) \mapsto \big(P(t),Q(t)\big)$, defined by (5.6), is the exact flow of the Hamiltonian system (1.7).

Proof. Whenever the mapping $(p,q) \mapsto \big(P(t),Q(t)\big)$ can be written as (5.12) with a function $S(q,Q,t)$, and when the invertibility assumption of Theorem 5.6 holds, the proof is done by the above calculations. Since our mapping, for $t=0$, reduces to the identity and cannot be written as (5.12), we give a direct proof.

Let $S^1(P,q,t)$ be given by the Hamilton–Jacobi equation (5.18), and assume that $(p,q) \mapsto (P,Q) = \big(P(t),Q(t)\big)$ is the transformation given by (5.6). Differentiation of the first relation of (5.6) with respect to time t and using (5.18) yields[5]

$$\Big(I + \frac{\partial^2 S^1}{\partial P \partial q}(P,q,t)\Big)\dot P \;=\; -\frac{\partial^2 S^1}{\partial t \partial q}(P,q,t) \;=\; -\Big(I + \frac{\partial^2 S^1}{\partial P \partial q}(P,q,t)\Big)\frac{\partial H}{\partial Q}(P,Q).$$

Differentiation of the second relation of (5.6) gives

$$\begin{aligned}\dot Q &= \frac{\partial^2 S^1}{\partial t \partial P}(P,q,t) + \frac{\partial^2 S^1}{\partial P^2}(P,q,t)\dot P \\ &= \frac{\partial H}{\partial P}(P,Q) + \frac{\partial^2 S^1}{\partial P^2}(P,q,t)\Big(\frac{\partial H}{\partial Q}(P,Q) + \dot P\Big).\end{aligned}$$

Consequently, $\dot P = -\frac{\partial H}{\partial Q}(P,Q)$ and $\dot Q = \frac{\partial H}{\partial P}(P,Q)$, so that $\big(P(t),Q(t)\big) = \varphi_t(p,q)$ is the exact flow of the Hamiltonian system. □

Writing the Hamilton–Jacobi differential equation in the variables $(P+p)/2$, $(Q+q)/2$ gives the following formula.

Theorem 5.8. *Assume that $S^3(u,v,t)$ is a solution of*

$$\frac{\partial S^3}{\partial t}(u,v,t) \;=\; H\Big(u - \frac{1}{2}\frac{\partial S^3}{\partial v}(u,v,t),\, v + \frac{1}{2}\frac{\partial S^3}{\partial u}(u,v,t)\Big) \tag{5.19}$$

with initial condition $S^3(u,v,0)=0$. Then, the exact flow $\varphi_t(p,q)$ of the Hamiltonian system (1.7) satisfies the system (5.7).

Proof. As in the proof of Theorem 5.7, one considers the transformation $(p,q) \mapsto \big(P(t),Q(t)\big)$ defined by (5.7), and then checks by differentiation that $\big(P(t),Q(t)\big)$ is a solution of the Hamiltonian system (1.7). □

Writing $w=(u,v)$ and using the matrix J of (2.3), the Hamilton–Jacobi equation (5.19) can also be written as

$$\frac{\partial S^3}{\partial t}(w,t) \;=\; H\Big(w + \frac{1}{2}J^{-1}\nabla S^3(w,t)\Big), \qquad S^3(w,0)=0. \tag{5.20}$$

The solution of (5.20) is anti-symmetric in t, i.e.,

$$S^3(w,-t) = -S^3(w,t). \tag{5.21}$$

[5] Due to an inconsistent notation of the partial derivatives $\frac{\partial H}{\partial Q}$, $\frac{\partial S^1}{\partial q}$ as column or row vectors, this formula may be difficult to read. Use indices instead of matrices in order to check its correctness.

This can be seen as follows: let $\varphi_t(w)$ be the exact flow of the Hamiltonian system $\dot y = J^{-1}\nabla H(y)$. Because of (5.8), $S^3(w,t)$ is defined by

$$\varphi_t(w) - w = J^{-1}\nabla S^3\big((\varphi_t(w)+w)/2, t\big).$$

Replacing t with $-t$ and then w with $\varphi_t(w)$ we get from $\varphi_{-t}\big(\varphi_t(t)\big) = w$ that

$$w - \varphi_t(w) = J^{-1}\nabla S^3\big((w+\varphi_t(w))/2, -t\big).$$

Hence $S^3(w,t)$ and $-S^3(w,-t)$ are generating functions of the same symplectic transformation. Since generating functions are unique up to an additive constant (because $dS = 0$ implies $S = Const$), the anti-symmetry (5.21) follows from the initial condition $S^3(w,0) = 0$.

VI.5.4 Methods Based on Generating Functions

To construct symplectic numerical methods of high order, Feng Kang (1986), Feng Kang, Wu, Qin & Wang (1989) and Channell & Scovel (1990) proposed computing an approximate solution of the Hamilton–Jacobi equation. For this one inserts the ansatz

$$S^1(P,q,t) = tG_1(P,q) + t^2G_2(P,q) + t^3G_3(P,q) + \ldots$$

into (5.18), and compares like powers of t. This yields

$$\begin{aligned} G_1(P,q) &= H(P,q),\\ G_2(P,q) &= \frac{1}{2}\Big(\frac{\partial H}{\partial P}\frac{\partial H}{\partial q}\Big)(P,q),\\ G_3(P,q) &= \frac{1}{6}\Big(\frac{\partial^2 H}{\partial P^2}\Big(\frac{\partial H}{\partial q}\Big)^2 + \frac{\partial^2 H}{\partial P\partial q}\frac{\partial H}{\partial P}\frac{\partial H}{\partial q} + \frac{\partial^2 H}{\partial q^2}\Big(\frac{\partial H}{\partial P}\Big)^2\Big)(P,q). \end{aligned}$$

If we use the truncated series

$$S^1(P,q) \;=\; hG_1(P,q) + h^2G_2(P,q) + \ldots + h^rG_r(P,q) \tag{5.22}$$

and insert it into (5.6), the transformation $(p,q)\mapsto(P,Q)$ defines a symplectic one-step method of order r. Symplecticity follows at once from Lemma 5.3 and order r is a consequence of the fact that the truncation of $S^1(P,q)$ introduces a perturbation of size $\mathcal{O}(h^{r+1})$ in (5.18). We remark that for $r\ge 2$ the methods obtained require the computation of higher derivatives of $H(p,q)$, and for separable Hamiltonians $H(p,q) = T(p) + U(q)$ they are no longer explicit (compared to the symplectic Euler method (3.1)).

The same approach applied to the third characterization of Lemma 5.3 yields

$$S^3(w,h) \;=\; hG_1(w) + h^3G_3(w) + \ldots + h^{2r-1}G_{2r-1}(w),$$

where $G_1(w) = H(w)$,

$$G_3(w) \;=\; \frac{1}{24}\,\nabla^2 H(w)\Big(J^{-1}\nabla H(w), J^{-1}\nabla H(w)\Big),$$

and further $G_j(w)$ can be obtained by comparing like powers of h in (5.20). In this way we get symplectic methods of order $2r$. Since $S^3(w,h)$ has an expansion in odd powers of h, the resulting method is symmetric.

The Approach of Miesbach & Pesch. With the aim of avoiding higher derivatives of the Hamiltonian in the numerical method, Miesbach & Pesch (1992) propose considering generating functions of the form

$$S^3(w,h) = h\sum_{i=1}^{s} b_i H\Big(w + hc_i J^{-1}\nabla H(w)\Big), \tag{5.23}$$

and to determine the free parameters b_i, c_i in such a way that the function of (5.23) agrees with the solution of the Hamilton–Jacobi equation (5.20) up to a certain order. For $b_{s+1-i} = b_i$ and $c_{s+1-i} = -c_i$ this function satisfies $S^3(w,-h) = -S^3(w,h)$, so that the resulting method is symmetric. A straightforward computation shows that it yields a method of order 4 if

$$\sum_{i=1}^{s} b_i = 1, \qquad \sum_{i=1}^{s} b_i c_i^2 = \frac{1}{12}.$$

For $s=3$, these equations are fulfilled for $b_1 = b_3 = 5/18$, $b_2 = 4/9$, $c_1 = -c_3 = \sqrt{15}/10$, and $c_2 = 0$. Since the function S^3 of (5.23) has to be inserted into (5.20), these methods still need second derivatives of the Hamiltonian.

VI.6 Variational Integrators

A third approach to symplectic integrators comes from using discretized versions of Hamilton's principle, which determines the equations of motion from a variational problem. This route has been taken by Suris (1990), MacKay (1992) and in a series of papers by Marsden and coauthors, see the review by Marsden & West (2001) and references therein. Basic theoretical properties were formulated by Maeda (1980,1982) and Veselov (1988,1991) in a non-numerical context.

VI.6.1 Hamilton's Principle

> Ours, according to Leibniz, is the best of all possible worlds, and the laws of nature can therefore be described in terms of extremal principles.
> (C.L. Siegel & J.K. Moser 1971, p. 1)

> Man scheint dies Princip früher ... unbemerkt gelassen zu haben. *Hamilton* ist der erste, der von diesem Princip ausgegangen ist.
> (C.G.J. Jacobi 1842, p. 58)

> Hamilton gave an improved mathematical formulation of a principle which was well established by the fundamental investigations of Euler and Lagrange; the integration process employed by him was likewise known to Lagrange. The name "Hamilton's principle", coined by Jacobi, was not adopted by the scientists of the last century. It came into use, however, through the textbooks of more recent date.
>
> (C. Lanczos 1949, p. 114)

Lagrange's equations of motion (1.4) can be viewed as the Euler–Lagrange equations for the variational problem of extremizing the *action integral*

$$\mathcal{S}(q) = \int_{t_0}^{t_1} L(q(t), \dot q(t))\, dt \tag{6.1}$$

among all curves $q(t)$ that connect two given points q_0 and q_1:

$$q(t_0) = q_0\,, \quad q(t_1) = q_1\,. \tag{6.2}$$

In fact, assuming $q(t)$ to be extremal and considering a variation $q(t) + \varepsilon\, \delta q(t)$ with the same end-points, i.e., with $\delta q(t_0) = \delta q(t_1) = 0$, gives, using a partial integration,

$$0 = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathcal{S}(q+\varepsilon\,\delta q) = \int_{t_0}^{t_1} \Big(\frac{\partial L}{\partial q}\,\delta q + \frac{\partial L}{\partial \dot q}\,\delta \dot q\Big) dt = \int_{t_0}^{t_1} \Big(\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot q}\Big) \delta q\, dt\,,$$

which leads to (1.4). The principle that the motion extremizes the action integral is known as *Hamilton's principle*.

We now consider the action integral as a function of (q_0, q_1), for the solution $q(t)$ of the Euler–Lagrange equations (1.4) with these boundary values (this exists uniquely locally at least if q_0, q_1 are sufficiently close),

$$S(q_0, q_1) = \int_{t_0}^{t_1} L(q(t), \dot q(t))\, dt\,. \tag{6.3}$$

The partial derivative of S with respect to q_0 is, again using partial integration,

$$\begin{aligned} \frac{\partial S}{\partial q_0} &= \int_{t_0}^{t_1} \Big(\frac{\partial L}{\partial q}\frac{\partial q}{\partial q_0} + \frac{\partial L}{\partial \dot q}\frac{\partial \dot q}{\partial q_0}\Big) dt \\ &= \frac{\partial L}{\partial \dot q}\frac{\partial q}{\partial q_0}\Big|_{t_0}^{t_1} + \int_{t_0}^{t_1} \Big(\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot q}\Big) \frac{\partial q}{\partial q_0}\, dt = -\frac{\partial L}{\partial \dot q}(q_0, \dot q_0) \end{aligned}$$

with $\dot q_0 = \dot q(t_0)$, where the last equality follows from (1.4) and (6.2). In view of the definition (1.5) of the conjugate momenta, $p = \partial L/\partial \dot q$, the last term is simply $-p_0$. Computing $\partial S/\partial q_1 = p_1$ in the same way, we thus obtain for the differential of S

$$dS = \frac{\partial S}{\partial q_1}\, dq_1 + \frac{\partial S}{\partial q_0}\, dq_0 = p_1\, dq_1 - p_0\, dq_0 \tag{6.4}$$

which is the basic formula for symplecticity generating functions (see (5.1) above), obtained here by working with the Lagrangian formalism.

VI.6.2 Discretization of Hamilton's Principle

Discrete-time versions of Hamilton's principle are of mathematical interest in their own right, see Maeda (1980,1982), Veselov (1991) and references therein. Here they are considered with the aim of deriving or understanding numerical approximation schemes. The discretized Hamilton principle consists of extremizing, for given q_0 and q_N, the sum

$$\mathcal{S}_h(\{q_n\}_0^N) = \sum_{n=0}^{N-1} L_h(q_n, q_{n+1}) \, . \tag{6.5}$$

We think of the *discrete Lagrangian* L_h as an approximation

$$L_h(q_n, q_{n+1}) \approx \int_{t_n}^{t_{n+1}} L(q(t), \dot{q}(t)) \, dt \, , \tag{6.6}$$

where $q(t)$ is the solution of the Euler–Lagrange equations (1.4) with boundary values $q(t_n) = q_n$, $q(t_{n+1}) = q_{n+1}$. If equality holds in (6.6), then it is clear from the continuous Hamilton principle that the exact solution values $\{q(t_n)\}$ of the Euler–Lagrange equations (1.4) extremize the action sum $\mathcal{S}_h$. Before we turn to concrete examples of approximations L_h, we continue with the general theory which is analogous to the continuous case.

The requirement $\partial \mathcal{S}_h / \partial q_n = 0$ for an extremum yields the *discrete Euler–Lagrange equations*

$$\frac{\partial L_h}{\partial y}(q_{n-1}, q_n) + \frac{\partial L_h}{\partial x}(q_n, q_{n+1}) = 0 \tag{6.7}$$

for $n = 1, \ldots, N-1$, where the partial derivatives refer to $L_h = L_h(x, y)$. This gives a three-term difference scheme for determining $q_1, \ldots, q_{N-1}$.

We now set

$$S_h(q_0, q_N) = \sum_{n=0}^{N-1} L_h(q_n, q_{n+1})$$

where $\{q_n\}$ is a solution of the discrete Euler–Lagrange equations (6.7) with the boundary values q_0 and q_N. With (6.7) the partial derivatives reduce to

$$\frac{\partial S_h}{\partial q_0} = \frac{\partial L_h}{\partial x}(q_0, q_1), \qquad \frac{\partial S_h}{\partial q_N} = \frac{\partial L_h}{\partial y}(q_{N-1}, q_N) \, .$$

We introduce the *discrete momenta* via a discrete Legendre transformation,

$$p_n = -\frac{\partial L_h}{\partial x}(q_n, q_{n+1}) \, . \tag{6.8}$$

The above formula and (6.7) for $n = N$ then yield

$$dS_h = p_N \, dq_N - p_0 \, dq_0 . \tag{6.9}$$

If (6.8) defines a bijection between p_n and q_{n+1} for given q_n, then we obtain a one-step method $\Phi_h : (p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$ by composing the inverse discrete Legendre transform, a step with the discrete Euler–Lagrange equations, and the discrete Legendre transformation as shown in the diagram:

$$\begin{array}{ccc} (q_n, q_{n+1}) & \overset{(6.7)}{\longrightarrow} & (q_{n+1}, q_{n+2}) \\ (6.8) \uparrow & & \downarrow (6.8) \\ (p_n, q_n) & & (p_{n+1}, q_{n+1}) \end{array}$$

The method is symplectic by (6.9) and Theorem 5.1. A short-cut in the computation is obtained by noting that (6.7) and (6.8) (for $n+1$ instead of n) imply

$$p_{n+1} = \frac{\partial L_h}{\partial y}(q_n, q_{n+1}) \,, \tag{6.10}$$

which yields the scheme

$$(p_n, q_n) \overset{(6.8)}{\longrightarrow} (q_n, q_{n+1}) \overset{(6.10)}{\longrightarrow} (p_{n+1}, q_{n+1}) \,.$$

Let us summarize these considerations, which can be found in Maeda (1980), Suris (1990), Veselov (1991) and MacKay (1992).

Theorem 6.1. *The discrete Hamilton principle for (6.5) gives the discrete Euler–Lagrange equations (6.7) and the symplectic method*

$$p_n = -\frac{\partial L_h}{\partial x}(q_n, q_{n+1}) \,, \qquad p_{n+1} = \frac{\partial L_h}{\partial y}(q_n, q_{n+1}) \,. \tag{6.11}$$

These formulas also show that L_h is a generating function (5.4) for the symplectic map $(p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$. Conversely, since every symplectic method has a generating function (5.4), it can be interpreted as resulting from Hamilton's principle with the generating function (5.4) as the discrete Lagrangian. The classes of symplectic integrators and variational integrators are therefore identical.

We now turn to simple examples of variational integrators obtained by choosing a discrete Lagrangian L_h with (6.6).

Example 6.2 (MacKay 1992). Choose $L_h(q_n, q_{n+1})$ by approximating $q(t)$ of (6.6) as the linear interpolant of q_n and q_{n+1} and approximating the integral by the trapezoidal rule. This gives

$$L_h(q_n, q_{n+1}) \;=\; \frac{h}{2}\, L\Big(q_n, \frac{q_{n+1}-q_n}{h}\Big) + \frac{h}{2}\, L\Big(q_{n+1}, \frac{q_{n+1}-q_n}{h}\Big) \tag{6.12}$$

and hence the symplectic scheme, with $v_{n+1/2} = (q_{n+1}-q_n)/h$ for brevity,

$$
\begin{aligned}
p_n &= \frac{1}{2}\frac{\partial L}{\partial \dot q}(q_n, v_{n+1/2}) + \frac{1}{2}\frac{\partial L}{\partial \dot q}(q_{n+1}, v_{n+1/2}) - \frac{h}{2}\frac{\partial L}{\partial q}(q_n, v_{n+1/2}) \\
p_{n+1} &= \frac{1}{2}\frac{\partial L}{\partial \dot q}(q_n, v_{n+1/2}) + \frac{1}{2}\frac{\partial L}{\partial \dot q}(q_{n+1}, v_{n+1/2}) + \frac{h}{2}\frac{\partial L}{\partial q}(q_{n+1}, v_{n+1/2}) \,.
\end{aligned}
$$

For a mechanical Lagrangian $L(q, \dot q) = \frac{1}{2}\dot q^T M \dot q - U(q)$ this reduces to the Störmer–Verlet method

$$
\begin{aligned}
M v_{n+1/2} &= p_n + \frac{1}{2} h F_n \\
q_{n+1} &= q_n + h v_{n+1/2} \\
p_{n+1} &= M v_{n+1/2} + \frac{1}{2} h F_{n+1}
\end{aligned}
$$

where $F_n = -\nabla U(q_n)$. In this case, the discrete Euler–Lagrange equations (6.7) become the familiar second-difference formula $M(q_{n+1} - 2q_n + q_{n-1}) = h^2 F_n$.

Example 6.3 (Wendlandt & Marsden 1997). Approximating the integral in (6.6) instead by the midpoint rule gives

$$
L_h(q_n, q_{n+1}) = hL\Big(\frac{q_{n+1} + q_n}{2}, \frac{q_{n+1} - q_n}{h}\Big). \tag{6.13}
$$

This yields the symplectic scheme, with the abbreviations $q_{n+1/2} = (q_{n+1} + q_n)/2$ and $v_{n+1/2} = (q_{n+1} - q_n)/h$,

$$
\begin{aligned}
p_n &= \frac{\partial L}{\partial \dot q}(q_{n+1/2}, v_{n+1/2}) - \frac{h}{2}\frac{\partial L}{\partial q}(q_{n+1/2}, v_{n+1/2}) \\
p_{n+1} &= \frac{\partial L}{\partial \dot q}(q_{n+1/2}, v_{n+1/2}) + \frac{h}{2}\frac{\partial L}{\partial q}(q_{n+1/2}, v_{n+1/2}) \,.
\end{aligned}
$$

For $L(q, \dot q) = \frac{1}{2}\dot q^T M \dot q - U(q)$ this becomes the implicit midpoint rule

$$
\begin{aligned}
M v_{n+1/2} &= p_n + \frac{1}{2} h F_{n+1/2} \\
q_{n+1} &= q_n + h v_{n+1/2} \\
p_{n+1} &= M v_{n+1/2} + \frac{1}{2} h F_{n+1/2}
\end{aligned}
$$

with $F_{n+1/2} = -\nabla U(\frac{1}{2}(q_{n+1} + q_n))$.

VI.6.3 Symplectic Partitioned Runge–Kutta Methods Revisited

To obtain higher-order variational integrators, Marsden & West (2001) consider the discrete Lagrangian

$$
L_h(q_0, q_1) = h \sum_{i=1}^{s} b_i L\big(u(c_i h), \dot u(c_i h)\big) \tag{6.14}
$$

where $u(t)$ is the polynomial of degree s with $u(0) = q_0$, $u(h) = q_1$ which extremizes the right-hand side. They then show that the corresponding variational integrator can be realized as a partitioned Runge–Kutta method. We here consider the slightly more general case

$$L_h(q_0, q_1) = h \sum_{i=1}^{s} b_i L(Q_i, \dot{Q}_i) \tag{6.15}$$

where

$$Q_i = q_0 + h \sum_{j=1}^{s} a_{ij} \dot{Q}_j$$

and the $\dot{Q}_i$ are chosen to extremize the above sum under the constraint

$$q_1 = q_0 + h \sum_{i=1}^{s} b_i \dot{Q}_i \ .$$

We assume that all the b_i are non-zero and that their sum equals 1. Note that (6.14) is the special case of (6.15) where the a_{ij} and b_i are integrals (II.1.10) of Lagrange polynomials as for collocation methods.

With a Lagrange multiplier $\lambda = (\lambda_1, \ldots, \lambda_d)$ for the constraint, the extremality conditions obtained by differentiating (6.15) with respect to $\dot{Q}_j$ for $j = 1, \ldots, s$, read

$$\sum_{i=1}^{s} b_i \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) h a_{ij} + b_j \frac{\partial L}{\partial \dot{q}}(Q_j, \dot{Q}_j) = b_j \lambda \ .$$

With the notation

$$\dot{P}_i = \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) \ , \qquad P_i = \frac{\partial L}{\partial \dot{q}}(Q_i, \dot{Q}_i) \tag{6.16}$$

this simplifies to

$$b_j P_j = b_j \lambda - h \sum_{i=1}^{s} b_i a_{ij} \dot{P}_i \ . \tag{6.17}$$

The symplectic method of Theorem 6.1 now becomes

$$\begin{aligned} p_0 &= -\frac{\partial L_h}{\partial x}(q_0, q_1) \\ &= -h \sum_{i=1}^{s} b_i \dot{P}_i \Big(I + h \sum_{j=1}^{s} a_{ij} \frac{\partial \dot{Q}_j}{\partial q_0} \Big) - h \sum_{j=1}^{s} b_j P_j \frac{\partial \dot{Q}_j}{\partial q_0} \\ &= -h \sum_{i=1}^{s} b_i \dot{P}_i + \lambda \ . \end{aligned}$$

In the last equality we use (6.17) and $h \sum_j b_j \partial \dot{Q}_j / \partial q_0 = -I$, which follows from differentiating the constraint. In the same way we obtain

$$p_1 = \frac{\partial L_h}{\partial y}(q_0, q_1) = \lambda \, .$$

Putting these formulas together, we see that (p_1, q_1) result from applying a partitioned Runge–Kutta method to the Lagrange equations (1.4) written as a differential-algebraic system

$$\dot{p} = \frac{\partial L}{\partial q}(q, \dot{q}) \, , \qquad p = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) \, . \tag{6.18}$$

That is

$$\begin{aligned} p_1 &= p_0 + h \sum_{i=1}^{s} b_i \dot{P}_i \, , \qquad & q_1 &= q_0 + h \textstyle\sum_{i=1}^{s} b_i \dot{Q}_i \, , \\ P_i &= p_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} \dot{P}_j \, , \qquad & Q_i &= q_0 + h \textstyle\sum_{j=1}^{s} a_{ij} \dot{Q}_j \, , \end{aligned} \tag{6.19}$$

with $\widehat{a}_{ij} = b_j - b_j a_{ji}/b_i$ so that the symplecticity condition (4.3) is fulfilled, and with $P_i, Q_i, \dot{P}_i, \dot{Q}_i$ related by (6.16). Since equations (6.16) are of the same form as (6.18), the proof of Theorem 1.3 shows that they are equivalent to

$$\dot{P}_i = -\frac{\partial H}{\partial q}(P_i, Q_i) \, , \qquad \dot{Q}_i = \frac{\partial H}{\partial p}(P_i, Q_i) \tag{6.20}$$

with the Hamiltonian $H = p^T \dot{q} - L(q, \dot{q})$ of (1.6). We have thus proved the following, which is similar in spirit to a result of Suris (1990).

Theorem 6.4. *The variational integrator with the discrete Lagrangian (6.15) is equivalent to the symplectic partitioned Runge–Kutta method (6.19), (6.20) applied to the Hamiltonian system with the Hamiltonian (1.6).* □

In particular, as noted by Marsden & West (2001), choosing Gaussian quadrature in (6.14) gives the Gauss collocation method applied to the Hamiltonian system, while Lobatto quadrature gives the Lobatto IIIA - IIIB pair.

VI.6.4 Noether's Theorem

> . . . enthält Satz I alle in Mechanik u.s.w. bekannten Sätze über erste Integrale. (E. Noether 1918)

We now return to the subject of Chap. IV, i.e., the existence of first integrals, but here in the context of Hamiltonian systems. E. Noether found the surprising result that continuous *symmetries* in the Lagrangian lead to such first integrals. We give in the following a version of her "Satz I", specialized to our needs, with a particularly short proof.

Theorem 6.5 (Noether 1918). *Consider a system with Hamiltonian $H(p, q)$ and Lagrangian $L(q, \dot{q})$. Suppose $\{g_s : s \in \mathbb{R}\}$ is a one-parameter group of transformations ($g_s \circ g_r = g_{s+r}$) which leaves the Lagrangian invariant:*

$$L(g_s(q), g_s'(q)\dot q) = L(q, \dot q) \quad \textit{for all } s \textit{ and all } (q, \dot q). \tag{6.21}$$

Let $a(q) = (d/ds)|_{s=0}\, g_s(q)$ *be defined as the vector field with flow* $g_s(q)$*. Then*

$$I(p, q) = p^T a(q) \tag{6.22}$$

is a first integral of the Hamiltonian system.

Example 6.6. Let G be a matrix Lie group with Lie algebra $\mathfrak{g}$ (see Sect. IV.6). Suppose $L(Qq, Q\dot q) = L(q, \dot q)$ for all $Q \in G$. Then $p^T Aq$ is a first integral for every $A \in \mathfrak{g}$. (Take $g_s(q) = \exp(sA)q$.) For example, $G = SO(n)$ yields conservation of angular momentum.

We prove Theorem 6.5 by using the discrete analogue, which reads as follows.

Theorem 6.7. *Suppose the one-parameter group of transformations* $\{g_s : s \in \mathbb{R}\}$ *leaves the discrete Lagrangian* $L_h(q_0, q_1)$ *invariant:*

$$L_h(g_s(q_0), g_s(q_1)) = L_h(q_0, q_1) \quad \textit{for all } s \textit{ and all } (q_0, q_1). \tag{6.23}$$

Then (6.22) is a first integral of the method (6.11), i.e., $p_{n+1}^T a(q_{n+1}) = p_n^T a(q_n)$*.*

Proof. Differentiating (6.23) with respect to s gives

$$0 = \frac{d}{ds}\Big|_{s=0} L_h(g_s(q_0), g_s(q_1)) = \frac{\partial L_h}{\partial x}(q_0, q_1)a(q_0) + \frac{\partial L_h}{\partial y}(q_0, q_1)a(q_1).$$

By (6.11) this becomes $0 = -p_0^T a(q_0) + p_1^T a(q_1)$. □

Theorem 6.5 now follows by choosing $L_h = S$ of (6.3) and noting (6.4) and

$$\begin{aligned} S(q(t_0), q(t_1)) &= \int_{t_0}^{t_1} L\big(q(t), \dot q(t)\big)\, dt \\ &= \int_{t_0}^{t_1} L\Big(g_s(q(t)), \frac{d}{dt} g_s(q(t))\Big)\, dt = S\Big(g_s(q(t_0)), g_s(q(t_1))\Big). \end{aligned}$$

Theorem 6.7 has the appearance of giving a rich source of first integrals for symplectic methods. However, it must be noted that, unlike the case of the exact flow map in the above formula, the invariance (6.21) of the Lagrangian L does not in general imply the invariance (6.23) of the discrete Lagrangian L_h of the numerical method. A noteworthy exception arises for linear transformations g_s as in Example 6.6, for which Theorem 6.7 yields the conservation of quadratic first integrals $p^T Aq$, such as angular momentum, by symplectic partitioned Runge–Kutta methods – a property we already know from Theorem IV.2.4. For Hamiltonian systems with an associated Lagrangian $L(q, \dot q) = \frac{1}{2}\dot q^T M \dot q - U(q)$, all first integrals originating from Noether's Theorem are quadratic (see Exercise 13).

VI.7 Characterization of Symplectic Methods

Up to now in this chapter, we have presented sufficient conditions for the symplecticity of numerical integrators (usually in terms of certain coefficients). Here, we will prove *necessary* conditions for symplecticity, i.e., answer the question as to which methods are *not* symplectic. It will turn out that the sufficient conditions of Sect. VI.4, under an irreducibility condition on the method, are also necessary. The main tool is the Taylor series expansion of the numerical flow $y_0 \mapsto \Phi_h(y_0)$, which we assume to be a B-series (or a P-series).

VI.7.1 B-Series Methods Conserving Quadratic First Integrals

The numerical solution of a Runge–Kutta method (II.1.4) can be written as a B-series

$$y_1 \;=\; B(a,y_0) \;=\; y_0 + \sum_{\tau\in T} \frac{h^{|\tau|}}{\sigma(\tau)}\, a(\tau)\, F(\tau)(y_0) \tag{7.1}$$

with coefficients $a(\tau)$ given by

$$a(\tau) = \sum_{i=1}^{s} b_i \mathbf{g}_i(\tau) \qquad \text{for} \quad \tau\in T \tag{7.2}$$

(see (III.1.16) and Sect. III.1.2). Our aim is to express the sufficient condition for the exact conservation of quadratic first integrals (which is the same as for symplecticity) in terms of the coefficients $a(\tau)$. For this we multiply (4.2) by $\mathbf{g}_i(u)\cdot\mathbf{g}_j(v)$ (where $u=[u_1,\ldots,u_m]$ and $v=[v_1,\ldots,v_l]$ are trees in T) and we sum over all i and j. Using (III.1.13) and the recursion (III.1.15) this yields

$$\sum_{i=1}^{s} b_i \mathbf{g}_i(u\circ v) + \sum_{j=1}^{s} b_j \mathbf{g}_j(v\circ u) \;=\; \Big(\sum_{i=1}^{s} b_i \mathbf{g}_i(u)\Big)\Big(\sum_{j=1}^{s} b_j \mathbf{g}_j(v)\Big),$$

where we have used the Butcher product (see, e.g., Butcher (1987), Sect. 143)

$$u\circ v = [u_1,\ldots,u_m,v], \qquad v\circ u = [v_1,\ldots,v_l,u] \tag{7.3}$$

(compare also Definition III.3.7 and Fig. 7.1 below). Because of (7.2), this implies

$$a(u\circ v) + a(v\circ u) = a(u)\cdot a(v) \qquad \text{for} \quad u,v\in T. \tag{7.4}$$

We now forget that the B-series (7.1) has been obtained from a Runge–Kutta method, and we ask the following question: is the condition (7.4) sufficient for a B-series method defined by (7.1) to conserve exactly quadratic first integrals (and to be symplectic)? The next theorem shows that this is indeed true, and we shall see later that condition (7.4) is also necessary (cf. Chartier, Faou & Murua 2005).

Theorem 7.1. *Consider a B-series method $\Phi_h(y) = B(a, y)$ and problems $\dot{y} = f(y)$ having $Q(y) = y^T Cy$ (with symmetric matrix C) as first integral.*

If the coefficients $a(\tau)$ satisfy (7.4), then the method exactly conserves $Q(y)$ and it is symplectic.

Proof. a) Under the assumptions of the theorem we shall prove in part (c) that

$$B(a,y)^T CB(a,y) = y^T Cy + \sum_{u,v\in T} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)} m(u,v)\, F(u)(y)^T CF(v)(y) \quad (7.5)$$

with $m(u,v) = a(u)\cdot a(v) - a(u\circ v) - a(v\circ u)$. Condition (7.4) is equivalent to $m(u,v) = 0$ and thus implies the exact conservation of $Q(y) = y^T Cy$.

To prove symplecticity of the method it is sufficient to show that the diagram of Lemma 4.1 commutes for general B-series methods. This is seen by differentiating the elementary differentials and by comparing them with those for the augmented system (Exercise 8). Symplecticity of the method thus follows as in Sect. VI.4.1 form the fact that the symplecticity relation is a quadratic first integral of the augmented system.

b) Since $Q(y) = y^T Cy$ is a first integral of $\dot{y} = f(y)$, we have $y^T Cf(y) = 0$ for all y. Differentiating m times this relation with respect to y yields

$$\sum_{j=1}^{m} k_j^T Cf^{(m-1)}(y)\big(k_1,\dots,k_{j-1},k_{j+1}\dots,k_m\big) + y^T Cf^{(m)}(y)\big(k_1,\dots,k_m\big) = 0.$$

Putting $k_j = F(\tau_j)(y)$ we obtain the formula

$$y^T CF([\tau_1,\dots,\tau_m])(y) = -\sum_{j=1}^{m} F(\tau_j)(y)^T CF([\tau_1,\dots,\tau_{j-1},\tau_{j+1},\dots,\tau_m])(y),$$

which can also be written in the form

$$y^T C\,\frac{F(\tau)(y)}{\sigma(\tau)} \;=\; -\sum_{u,v\in T, v\circ u=\tau} \frac{F(u)(y)^T}{\sigma(u)}\, C\, \frac{F(v)(y)}{\sigma(v)}. \quad (7.6)$$

c) With (7.1) the expression $y_1^T Cy_1$ becomes

$$\begin{aligned} B(a,y)^T CB(a,y) \;&=\; y^T Cy + 2y^T C \sum_{\tau\in T} \frac{h^{|\tau|}}{\sigma(\tau)}\, a(\tau)\, F(\tau)(y) \\ &+\; \sum_{u,v\in T} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)}\, a(u)\, a(v)\, F(u)(y)^T CF(v)(y). \end{aligned}$$

Since C is symmetric, formula (7.6) remains true if we sum over trees u, v such that $u\circ v = \tau$. Inserting both formulas into the sum over τ leads directly to (7.5). □

Extension to P-Series. All the previous results can be extended to partitioned methods. To find the correct conditions on the coefficients of the P-series, we use the fact that the numerical solution of a partitioned Runge–Kutta method (II.2.2) is a P-series

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} P_p(a,(p_0,q_0)) \\ P_q(a,(p_0,q_0)) \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + \begin{pmatrix} \sum_{u\in TP_p} \frac{h^{|u|}}{\sigma(u)}\, a(u)\, F(u)(p_0,q_0) \\ \sum_{v\in TP_q} \frac{h^{|v|}}{\sigma(v)}\, a(v)\, F(v)(p_0,q_0) \end{pmatrix} \tag{7.7}$$

with coefficients $a(\tau)$ given by

$$a(\tau) = \begin{cases} \sum_{i=1}^s b_i \phi_i(\tau) & \text{for} \quad \tau \in TP_p \\ \sum_{i=1}^s \widehat{b}_i \phi_i(\tau) & \text{for} \quad \tau \in TP_q \end{cases} \tag{7.8}$$

(see Theorem III.2.4). We assume here that the elementary differentials $F(\tau)(p,q)$ originate from a partitioned sytem

$$\dot p = f_1(p,q), \quad \dot q = f_2(p,q), \tag{7.9}$$

such as the Hamiltonian system (1.7). This time we multiply (4.3) by $\phi_i(u)\cdot\phi_j(v)$ (where $u = [u_1,\ldots,u_m]_p \in TP_p$ and $v = [v_1,\ldots,v_l]_q \in TP_q$) and we sum over all i and j. Using the recursion (III.2.7) this yields

$$\sum_{i=1}^s b_i\phi_i(u\circ v) + \sum_{j=1}^s \widehat{b}_j\phi_j(v\circ u) = \Big(\sum_{i=1}^s b_i\phi_i(u)\Big)\Big(\sum_{j=1}^s \widehat{b}_j\phi_j(v)\Big), \tag{7.10}$$

where $u\circ v = [u_1,\ldots,u_m,v]_p$ and $v\circ u = [v_1,\ldots,v_l,u]_q$. Because of (7.8), this implies the relation

$$a(u\circ v) + a(v\circ u) = a(u)\cdot a(v) \qquad \text{for} \quad u\in TP_p,\ v\in TP_q. \tag{7.11}$$

Since $\phi_i(\tau)$ is independent of the colour of the root of τ, condition (4.4) implies

$$a(\tau) \ \text{ is independent of the colour of the root of } \tau. \tag{7.12}$$

Theorem 7.2. *Consider a P-series method $(p_1,q_1) = \Phi_h(p_0,q_0)$ given by (7.7), and a problem (7.9) having $Q(p,q) = p^T E\, q$ as first integral.*

i) If the coefficients $a(\tau)$ satisfy (7.11) and (7.12), the method exactly conserves $Q(p,q)$ and it is symplectic for general Hamiltonian systems (1.7).

ii) If the coefficients $a(\tau)$ satisfy only (7.11), the method exactly conserves $Q(p,q)$ for problems of the form $\dot p = f_1(q)$, $\dot q = f_2(p)$, and it is symplectic for separable Hamiltonian systems where $H(p,q) = T(p) + U(q)$.

Proof. This is very similar to that of Theorem 7.1. If $Q(p,q) = p^T E\, q$ is a first integral of (7.9), we have $f_1(p,q)^T E\, q + p^T E\, f_2(p,q) = 0$ for all p and q. Differentiating m times with respect to p and n times with respect to q yields

$$
\begin{aligned}
0 \;=\;& D_p^m D_q^n f_1(p,q)\big(k_1,\ldots,k_m,\ell_1,\ldots,\ell_n\big)^T E\, q \\
&+\; p^T E\, D_p^m D_q^n f_2(p,q)\big(k_1,\ldots,k_m,\ell_1,\ldots,\ell_n\big) \qquad (7.13)\\
&+\; \sum_{j=1}^{n} D_p^m D_q^{n-1} f_1(p,q)\big(k_1,\ldots,k_m,\ell_1,\ldots,\ell_{j-1},\ell_{j+1},\ldots,\ell_n\big)^T E\, \ell_j \\
&+\; \sum_{i=1}^{m} k_i^T E\, D_p^{m-1} D_q^n f_2(p,q)\big(k_1,\ldots,k_{i-1},k_{i+1},\ldots,k_m,\ell_1,\ldots,\ell_n\big).
\end{aligned}
$$

Putting $k_i = F(u_i)(p,q)$ with $u_i \in TP_p$, $\ell_j = F(v_j)(p,q)$ with $v_j \in TP_q$, $\tau_p = [u_1,\ldots,u_m,v_1,\ldots,v_n]_p$ and $\tau_q = [u_1,\ldots,u_m,v_1,\ldots,v_n]_q$, we obtain as in part (b) of the proof of Theorem 7.1 that

$$
\begin{aligned}
&\frac{F(\tau_p)(p,q)^T}{\sigma(\tau_p)} E\, q + p^T E\, \frac{F(\tau_q)(p,q)}{\sigma(\tau_q)} \qquad (7.14)\\
&= \sum_{u\circ v=\tau_p} \frac{F(u)(p,q)^T}{\sigma(u)} E\, \frac{F(v)(p,q)}{\sigma(v)} + \sum_{v\circ u=\tau_q} \frac{F(u)(p,q)^T}{\sigma(u)} E\, \frac{F(v)(p,q)}{\sigma(v)},
\end{aligned}
$$

where the sums are over $u \in TP_p$ and $v \in TP_q$.

With (7.7) the expression $p_1^T E\, q_1$ becomes

$$
\begin{aligned}
&P_p(a,(p,q))^T E\, P_q(a,(p,q)) \;=\; p^T E\, q \qquad (7.15)\\
&+ \sum_{u\in TP_p} \frac{h^{|u|}}{\sigma(u)}\, a(u)\, F(u)(p,q)^T E\, q \;+\; p^T E \sum_{v\in TP_q} \frac{h^{|v|}}{\sigma(v)}\, a(v)\, F(v)(p,q) \\
&+ \sum_{u\in TP_p, v\in TP_q} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)}\, a(u)a(v)\, F(u)(p,q)^T E\, F(v)(p,q).
\end{aligned}
$$

Condition (7.12) implies that $a(\tau_p) = a(\tau_q)$ for the trees in (7.14). Since also $|\tau_p| = |\tau_q|$ and $\sigma(\tau_p) = \sigma(\tau_q)$, two corresponding terms in the sums of the second line in (7.15) can be jointly replaced by the use of (7.14). As in part (c) of the proof of Theorem 7.1 this together with (7.11) then yields

$$
P_p(a,(p,q))^T E\, P_q(a,(p,q)) = p^T E\, q,
$$

which proves the conservation of quadratic first integrals $p^T E\, q$. Symplecticity follows as before, because the diagram of Lemma 4.1 also commutes for general P-series methods.

For the proof of statement (ii) we notice that $f_1(q)^T E\, q + p^T E\, f_2(p) = 0$ implies that $f_1(q)^T E\, q = 0$ and $p^T E\, f_2(p) = 0$ vanish separately. Instead of (7.14) we thus have two identities: the term $F(\tau_p)(p,q)^T E\, q/\sigma(\tau_p)$ becomes equal to the first sum in (7.14), and $p^T E\, F(\tau_q)(p,q)/\sigma(\tau_q)$ to the second sum. Consequently, the previous argumentation can be applied without the condition $a(\tau_p) = a(\tau_q)$. □

Second Order Differential Equations. We next consider partitioned systems of the particular form

$$\dot p = f_1(q), \qquad \dot q = Cp + c, \tag{7.16}$$

where C is a matrix and c a vector. Since problems of this type are second order differential equations $\ddot q = Cf_1(q)$, partitioned Runge–Kutta methods become equivalent to Nyström methods (see Sects. II.2.3 and IV.2.3).

An important special case are Hamiltonian systems

$$\dot p = -\nabla U(q), \qquad \dot q = Cp + c \tag{7.17}$$

(or, equivalently, $\ddot q = -C\nabla U(q)$). They correspond to Hamiltonian functions

$$H(p,q) = \frac{1}{2}\, p^T C p + c^T p + U(q), \tag{7.18}$$

where the kinetic energy is at most quadratic in p (here, C is usually symmetric).

In a P-series representation of the numerical solution, many elementary differentials vanish identically. Only those trees have to be considered, whose neighbouring vertices have different colour (the problem is separable) and whose white vertices have at most one son[6] (second component is linear). We denote this set of trees by

$$TN_p = \Big\{ \tau \in TP_p \;\Big|\; \begin{array}{c} \text{neighbouring vertices of } \tau \text{ have different colour} \\ \text{white vertices of } \tau \text{ have at most one son} \end{array} \Big\}, \tag{7.19}$$

and we let TN_q be the corresponding subset of TP_q.

The same procedure as for partitioned methods permits us to write the symplecticity condition of Theorem 4.8 in terms of the coefficients $a(\tau)$ of the P-series. Assuming $a(\bullet) = a(\circ) = 1$, the two conditions of (4.5) lead to

$$a(\circ \circ u) + a(u \circ \circ) = a(u)\, a(\circ) \qquad \text{for } u \in TN_p \tag{7.20}$$

$$a(u)a(\circ \circ v) - a(u \circ\circ v) = a(\circ \circ u)a(v) - a(v \circ\circ u) \quad \text{for } u, v \in TN_p \tag{7.21}$$

where we use the abbreviating notation

$$u \circ\circ v = u \circ (\circ \circ v) = [u_1, \ldots, u_m, [v]_q]_p \tag{7.22}$$

if $u = [u_1, \ldots, u_m]_p$. Notice that for $u, v \in TN_p$, the trees $u \circ \circ$, $u \circ\circ v$ and $v \circ\circ u$ are in TN_p, and $\circ \circ u$ is in TN_q.

Theorem 7.3. *Consider a P-series method (7.7) for differential equations (7.16) having $Q(p,q) = p^T E q$ as first integral.*

If the coefficients $a(\tau)$ satisfy (7.20) and (7.21), the method exactly conserves $Q(p,q)$ and it is symplectic for Hamiltonian systems with $H(p,q)$ of the form (7.18).

[6] Attention: with respect to (III.2.10) the vertices have opposite colour, because the linear dependence is in the second component in (7.17) whereas it is in the first component in (III.2.9).

Proof. Since the elementary differentials $F(\tau)(p,q)$ vanish identically for $\tau \notin TN_p \cup TN_q$, we can arbitrarily define $a(\tau)$ for trees outside $TN_p \cup TN_q$ without changing the method (throughout this proof we implicitly assume that for the considered trees neighbouring vertices have different colour). We shall do this in such a way that (7.11) holds.

Consider first the tree $u \circ\circ v$. There is exactly one vertex between the roots of u and v. Making this vertex to the root gives the tree $[u,v]_q$ which is not in TN_q. We define for $u, v \in TN_p$

$$a([u,v]_q) := a(u)a(\circ \circ v) - a(u \circ\circ v).$$

Condition (7.21) shows that $a([u,v]_q)$ is independent of permuting u and v and is thus well-defined. For trees that are neither in $TN_p \cup TN_q$ nor of the form $[u,v]_q$ with $u, v \in TN_p$ we let $a(\tau) = 0$. This extension of $a(\tau)$ implies that condition (7.11) holds for all trees, and part (ii) of Theorem 7.2 yields the statement. Notice that for problems $\dot p = f_1(q)$, $\dot q = f_2(p)$ only trees, for which neighbouring vertices have different colour, are relevant. □

VI.7.2 Characterization of Symplectic P-Series (and B-Series)

A characterization of symplectic B-series was first obtained by Calvo & Sanz-Serna (1994). We also consider P-series with various important special situations.

Theorem 7.4. *Consider a P-series method (7.7) applied to a general partitioned differential equation (7.9). Equivalent are:*

1) *the coefficients $a(\tau)$ satisfy (7.11) and (7.12),*
2) *quadratic first integrals of the form $Q(p,q) = p^T E\, q$ are exactly conserved,*
3) *the method is symplectic for general Hamiltonian systems (1.7).*

Proof. The implication (1)⇒(2) follows from part (i) of Theorem 7.2, (2)⇒(3) is a consequence of the fact that the symplecticity condition is a quadratic first integral of the variational equation (see the proof of Theorem 7.2). The remaining implication (3)⇒(1) will be proved in the following two steps.

a) We fix two trees $u \in TP_p$ and $v \in TP_q$, and we construct a (polynomial) Hamiltonian such that the transformation (7.7) satisfies

$$\Big(\frac{\partial(p_1,q_1)}{\partial p_0^1}\Big)^T J\,\Big(\frac{\partial(p_1,q_1)}{\partial q_0^2}\Big) \;=\; C\Big(a(u\circ v) + a(v\circ u) - a(u)\cdot a(v)\Big) \qquad (7.23)$$

with $C \neq 0$ (here, p_0^1 denotes the first component of p_0, and q_0^2 the second component of q_0). The symplecticity of (7.7) implies that the expression in (7.23) vanishes, so that condition (7.11) has to be satisfied.

For given $u \in TP_p$ and $v \in TP_q$ we define the Hamiltonian as follows: to the branches of $u \circ v$ we attach the numbers $3, \ldots, |u| + |v| + 1$ such that the branch between the roots of u and v is labelled by 3. Then, the Hamiltonian is a sum of as many terms as vertices in the tree. The summand corresponding to a vertex is a

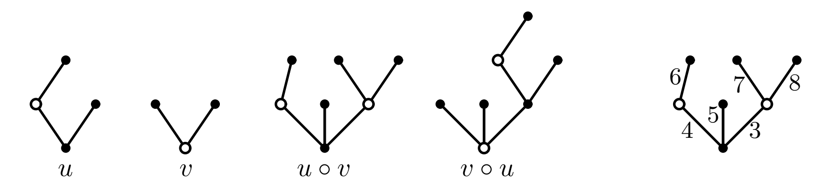


Fig. 7.1. Illustration of the Hamiltonian (7.24)

product containing the factor p^j (resp. q^j) if an upward leaving branch "j" is directly connected with a black (resp. white) vertex, and the factor q^i (resp. p^i) if the vertex itself is black (resp. white) and the downward leaving branch has label "i". Finally, the factors q^2 and p^1 are included in the terms corresponding to the roots of u and v, respectively. For the example of Fig. 7.1 we have

$$H(p,q) = q^2q^3q^4p^5 + p^1p^3p^7p^8 + p^4p^6 + q^5 + q^6 + q^7 + q^8. \qquad (7.24)$$

The components $F^i(\tau)(p,q)$ of the elementary differentials corresponding to the Hamiltonian system (with the Hamiltonian constructed above) satisfy

$$\begin{aligned} F^2(u\circ v)(p,q) &= (-1)^{\delta(u\circ v)}\sigma(u\circ v)\cdot p^1, \\ F^1(v\circ u)(p,q) &= (-1)^{\delta(v\circ u)}\sigma(v\circ u)\cdot q^2, \\ F^3(u)(p,q) &= (-1)^{\delta(u)}\sigma(u)\cdot q^2 \\ F^3(v)(p,q) &= (-1)^{\delta(v)}\sigma(v)\cdot p^1, \end{aligned} \qquad (7.25)$$

and for all other trees $\tau \in TP$ and components i we have

$$\frac{\partial F^i(\tau)}{\partial p^1}(0,0) = \frac{\partial F^i(\tau)}{\partial q^2}(0,0) = 0.$$

In (7.25), $\delta(\tau)$ counts the number of black vertices of τ, and the *symmetry coefficient* $\sigma(\tau)$ is that of (III.2.3). For example, $\sigma(u) = 1$ and $\sigma(v) = 2$ for the trees of Fig. 7.1. The verification of (7.25) is straightforward. The coefficient $(-1)^{\delta(\tau)}$ is due to the minus sign in the first part of the Hamiltonian system (1.7), and the symmetry coefficient $\sigma(\tau)$ appears in exactly the same way as in the multidimensional Taylor formula. Due to the zero initial values, no elementary differential other than those of (7.25) give rise to non-vanishing expressions in (7.23). Consider for example the second component of $F(\tau)(p,q)$ for a tree $\tau \in TP_p$. Since we are concerned with the Hamiltonian system (1.7), this expression starts with a derivative of H_{q^2}. Therefore, it contributes to (7.23) at $p_0 = q_0 = 0$ only if it contains the factor $H_{q^2q^3q^4p^5}$ (for the example of Fig. 7.1). This in turn implies the presence of factors $H_{p^3\ldots}$, $H_{p^4\ldots}$ and $H_{q^5\ldots}$. Continuing this line of reasoning, we find that $F^2(\tau)(p,q)$ contributes to (7.23) at $p_0 = q_0 = 0$ only if $\tau = u\circ v$. With similar arguments we see that only the elementary differentials of (7.25) have to be considered. We now insert (7.25) into (7.7), and we compute its derivatives with respect to p^1 and q^2. This then yields (7.23) with $C = (-1)^{\delta(u)+\delta(v)}h^{|u|+|v|}$, and completes the proof concerning condition (7.11).

b) The necessity of condition (7.12) is seen similarly. We fix a tree $\tau \in TP_p$ and we let $\overline{\tau} \in TP_q$ be the tree obtained from τ by changing the colour of the root. We then attach the numbers $3, \ldots, |\tau| + 1$ to the branches of τ, and we define a Hamiltonian as above but, different from adding the factors q^2 and p^1, we include the factor $p^1 q^2$ to the term corresponding to the root. For the tree $\tau = u$ of Fig. 7.1 this yields

$$H(p,q) \;=\; p^1 q^2 q^3 p^4 + p^3 p^5 + q^4 + q^5.$$

With this Hamiltonian we get

$$\begin{aligned} F^2(\tau)(p,q) &= (-1)^{\delta(\tau)}\sigma(\tau)\cdot p^1, \\ F^1(\overline{\tau})(p,q) &= (-1)^{\delta(\tau)}\sigma(\tau)\cdot q^2, \end{aligned}$$

and these are the only elementary differentials contributing to the left-hand expression of (7.23). We thus get

$$\Big(\frac{\partial(p_1,q_1)}{\partial p_0^1}\Big)^T J \,\Big(\frac{\partial(p_1,q_1)}{\partial q_0^2}\Big) \;=\; (-1)^{\delta(\tau)} h^{|\tau|}\Big(a(\tau) - a(\overline{\tau})\Big),$$

which completes the proof of Theorem 7.4. □

Theorem 7.5. *Consider a P-series method (7.7) applied to a separable partitioned differential equation $\dot p = f_1(q)$, $\dot q = f_2(p)$. Equivalent are:*
1) the coefficients $a(\tau)$ satisfy (7.11),
2) quadratic first integrals of the form $Q(p,q) = p^T E\, q$ are exactly conserved,
3) the method is symplectic for separable Hamiltonians $H(p,q) = T(p) + U(q)$.

Proof. The implications (1)⇒(2)⇒(3) follow as before from part (ii) of Theorem 7.2. The remaining implication (3)⇒(1) is a consequence of the fact that the Hamiltonian constructed in part (a) of the proof of Theorem 7.4 is separable, when u and v have no neighbouring vertices of the same colour. □

Theorem 7.6. *Consider a B-series method (7.1) for $\dot y = f(y)$. Equivalent are:*
1) the coefficients $a(\tau)$ satisfy (7.4),
2) quadratic first integrals of the form $Q(y) = y^T C y$ are exactly conserved,
3) the method is symplectic for general Hamiltonian systems $\dot y = J^{-1}\nabla H(y)$.

Proof. The implications (1)⇒(2)⇒(3) follow from Theorem 7.1. The remaining implication (3)⇒(1) follows from Theorem 7.4, because a B-series with coefficients $a(\tau), \tau \in T$, applied to a partitioned differential equation, can always be interpreted as a P-series (Definition III.2.1), where $a(\tau) := a(\varphi(\tau))$ for $\tau \in TP$ and $\varphi : TP \to T$ is the mapping that forgets the colouring of the vertices. This follows from the fact that

$$\alpha(\tau) F(\tau)(y) \;=\; \begin{pmatrix} \sum_{u\in TP_p,\varphi(u)=\tau} \alpha(u)\, F(u)(p,q) \\ \sum_{v\in TP_q,\varphi(v)=\tau} \alpha(v)\, F(v)(p,q) \end{pmatrix}$$

for $\tau \in T$, because $\alpha(u)\cdot\sigma(u) = \alpha(v)\cdot\sigma(v) = \mathbf{e}(\tau)\cdot|\tau|!\,$. Here, $y = (p,q)$, the elementary differentials $F(\tau)(y)$ are those of Definition III.1.2, whereas $F(u)(p,q)$ and $F(v)(p,q)$ are those of Table III.2.1. □

Theorem 7.7. *Consider a P-series method (7.7) applied to the special partitioned system (7.16). Equivalent are:*

1) the coefficients $a(\tau)$ satisfy (7.20) and (7.21),

2) quadratic first integrals of the form $Q(p,q) = p^T E\, q$ are exactly conserved,

3) the method is symplectic for Hamiltonian systems of the form (7.17).

Proof. The implications (1)⇒(2)⇒(3) follow from Theorem 7.3. The remaining implication (3)⇒(1) can be seen as follows.

Condition (7.20) is a consequence of the the proof of Theorem 7.4, because for $u \in TN_p$ and $v = \circ$ the Hamiltonian constructed there is of the form (7.18).

To prove condition (7.21) we have to modify slightly the definition of $H(p,q)$. We take $u, v \in TN_p$ and define the polynomial Hamiltonian as follows: to the branches of $u \circ\circ v$ we attach the numbers $3, \ldots, |u| + |v| + 2$. The Hamiltonian is then a sum of as many terms as vertices in the tree. The summands are defined as in the proof of Theorem 7.4 with the only exception that to the terms corresponding to the roots of u and v we include the factors q^2 and q^1, respectively, instead of q^2 and p^1. This gives a Hamiltonian of the form (7.18), for which the expression

$$\Big(\frac{\partial(p_1,q_1)}{\partial q_0^1}\Big)^T J \Big(\frac{\partial(p_1,q_1)}{\partial q_0^2}\Big) \tag{7.26}$$

becomes equal to

$$a(u)a(\circ\circ v) - a(u \circ\circ v) - a(\circ\circ u)a(v) + a(v \circ\circ u) \tag{7.27}$$

up to a nonzero constant. By symplecticity, (7.26) is zero so that also (7.27) has to vanish. This proves the validity of condition (7.21). □

VI.7.3 Irreducible Runge–Kutta Methods

We are now able to study to what extent the conditions of Theorem 4.3 and Theorem 4.6 are also necessary for symplecticity. Consider first the 2-stage method

$$\begin{array}{c|cc} 1/2 & \alpha & 1/2-\alpha \\ 1/2 & \beta & 1/2-\beta \\ \hline & 1/2 & 1/2 \end{array} \,.$$

The solution of the corresponding Runge–Kutta system (II.1.4) is given by $k_1 = k_2 = k$, where $k = f(y_0 + k/2)$, and hence $y_1 = y_0 + hk$. Whatever the values of α and β are, the numerical solution of the Runge–Kutta method is identical to that of the implicit midpoint rule, so that it defines a symplectic transformation. However, the condition (4.2) is only satisfied for $\alpha = \beta = 1/4$.

Definition 7.8. Two stages i and j of a Runge–Kutta method (II.1.4) are said to be *equivalent for a class $(\mathcal{P})$* of initial value problems, if for every problem in $(\mathcal{P})$ and for every sufficiently small step size we have $k_i = k_j$ ($k_i = k_j$ and $\ell_i = \ell_j$ for partitioned Runge–Kutta methods (II.2.2)).

The method is called *irreducible for ($\mathcal{P}$)* if it does not have equivalent stages. It is called *irreducible* if it is irreducible for all sufficiently smooth initial value problems.

For a more amenable characterization of irreducible Runge–Kutta methods, we introduce an ordering on T (and on TP), and we consider the following $s \times \infty$ matrices

$\Phi_{\text{RK}} = \big(\phi(\tau); \tau \in T\big)$ with entries $\phi_i(\tau) = \mathbf{g}_i(\tau)$ given by (III.1.13),[7]
$\Phi_{\text{PRK}} = \big(\phi(\tau); \tau \in TP_p\big) = \big(\phi(\tau); \tau \in TP_q\big)$ with entries $\phi_i(\tau)$ given by (III.2.7); observe that $\phi_i(\tau)$ does not depend on the colour of the root,
$\Phi^*_{\text{PRK}} = \big(\phi(\tau); \tau \in TP^*_p\big) = \big(\phi(\tau); \tau \in TP^*_q\big)$ where TP^*_p (resp. TP^*_q) is the set of trees in TP_p (resp. TP_q) whose neighbouring vertices have different colours.

Lemma 7.9 (Hairer 1994). *A Runge–Kutta method is irreducible if and only if the matrix Φ_{RK} has full rank s.*

A partitioned Runge–Kutta method is irreducible if and only if the matrix Φ_{PRK} has full rank s.

A partitioned Runge–Kutta method is irreducible for separable problems $\dot p = f_1(q)$, $\dot q = f_2(p)$ if and only if the matrix Φ^*_{PRK} has full rank s.

Proof. If the stages i and j are equivalent, it follows from the expansion

$$k_i = \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)}\, \phi_i(\tau)\, F(\tau)(y_0)$$

(see the proof of Theorem III.1.4) and from the independency of the elementary differentials (Exercise III.3) that $\phi_i(\tau) = \phi_j(\tau)$ for all $\tau \in T$. Hence, the rows i and j of the matrix Φ_{RK} are identical. The analogous statement for partitioned Runge–Kutta methods follows from Theorem III.2.4 and Exercise III.6. This proves the sufficiency of the "full rank" condition.

We prove its necessity only for partitioned Runge–Kutta methods applied to separable problems (the other situations can be treated similarly). For separable problems, only trees in $TP^*_p \cup TP^*_q$ give rise to non-vanishing elementary differentials. Irreducibility therefore implies that for every pair (i,j) with $i \neq j$ there exists a tree $\tau \in TP^*_p$ such that $\phi_i(\tau) \neq \phi_j(\tau)$. Consequently, a certain finite linear combination of the columns of Φ^*_{PRK} has distinct elements, i.e., there exist vectors $\xi \in \mathbb{R}^\infty$ (only finitely many non zero elements) and $\eta \in \mathbb{R}^s$ with $\Phi^*_{\text{PRK}}\xi = \eta$ and $\eta_i \neq \eta_j$ for $i \neq j$. Due to the fact that $\phi_i([\tau_1,\ldots,\tau_m]) = \phi_i([\tau_1]) \cdot \ldots \cdot \phi_i([\tau_m])$, the componentwise product of two columns of Φ^*_{PRK} is again a column of Φ^*_{PRK}. Continuing this argumentation and observing that $(1,\ldots,1)^T$ is a column of Φ^*_{PRK}, we obtain a matrix X such that $\Phi^*_{\text{PRK}}X = (\eta_i^{j-1})_{i,j=1}^s$ is a Vandermonde matrix. Since the η_i are distinct, the matrix Φ^*_{PRK} has to be of full rank s. □

[7] In this section we let $\phi(\tau) \in \mathbb{R}^s$ denote the vector whose elements are $\phi_i(\tau)$, $i = 1,\ldots,s$. This should not be mixed up with the value $\phi(\tau)$ of (III.1.16).

VI.7.4 Characterization of Irreducible Symplectic Methods

The necessity of the condition (4.2) for symplectic Runge–Kutta methods was first stated by Lasagni (1988). Abia & Sanz-Serna (1993) extended his proof to partitioned methods. We follow here the ideas of Hairer (1994).

Theorem 7.10. *An irreducible Runge–Kutta method (II.1.4) is symplectic if and only if the condition (4.2) holds.*

An irreducible partitioned Runge–Kutta method (II.2.2) is symplectic if and only if the conditions (4.3) and (4.4) hold.

A partitioned Runge–Kutta method, irreducible for separable problems, is symplectic for separable Hamiltonians $H(p,q) = T(p) + U(q)$ if and only if the condition (4.3) holds.

Proof. The "if" part of all three statements has been proved in Theorem 4.3 and Theorem 4.6. We prove the "only if" part for partitioned Runge–Kutta methods applied to general Hamiltonian systems (the other two statements can be obtained in the same way).

We consider the $s \times s$ matrix M with entries $m_{ij} = b_i \widehat{a}_{ij} + \widehat{b}_j a_{ji} - b_i \widehat{b}_j$. The computation leading to formula (7.11) shows that for $u \in TP_p$ and $v \in TP_q$

$$\phi(u)^T M\, \phi(v) \;=\; a(u \circ v) + a(v \circ u) - a(u) \cdot a(v)$$

holds. Due to the symplecticity of the method, this expression vanishes and we obtain

$$\Phi_{\text{PRK}}^T M\, \Phi_{\text{PRK}} \;=\; 0,$$

where Φ_{PRK} is the matrix of Lemma 7.9. An application of this lemma then yields $M = 0$, which proves the necessity of (4.3).

For the vector d with components $d_i = b_i - \widehat{b}_i$ we get $d^T \Phi_{\text{PRK}} = 0$, and we deduce from Lemma 7.9 that $d = 0$, so that (4.4) is also seen to be necessary. □

VI.8 Conjugate Symplecticity

The symplecticity requirement may be too strong if we are interested in a correct long-time behaviour of a numerical integrator. Stoffer (1988) suggests considering methods that are not necessarily symplectic but conjugate to a symplectic method.

Definition 8.1. Two numerical methods Φ_h and Ψ_h are mutually *conjugate*, if there exists a global change of coordinates χ_h, such that

$$\Phi_h = \chi_h^{-1} \circ \Psi_h \circ \chi_h. \tag{8.1}$$

We assume that $\chi_h(y) = y + \mathcal{O}(h)$ uniformly for y varying in a compact set.

For a numerical solution $y_{n+1} = \Phi_h(y_n)$, lying in a compact subset of the phase space, the transformed values $z_n = \chi_h(y_n)$ constitute a numerical solution $z_{n+1} = \Psi_h(z_n)$ of the second method. Since $y_n - z_n = \mathcal{O}(h)$, both numerical solutions have the same long-time behaviour, independently of whether one method shares certain properties (e.g., symplecticity) with the other.

VI.8.1 Examples and Order Conditions

The most prominent pair of conjugate methods are the trapezoidal and midpoint rules. Their conjugacy has been originally exploited by Dahlquist (1975) in an investigation on nonlinear stability.

If we denote by Φ_h^E and Φ_h^I the explicit and implicit Euler methods, respectively, then the trapezoidal rule Φ_h^T and the implicit midpoint rule Φ_h^M can be written as

$$\Phi_h^T = \Phi_{h/2}^I \circ \Phi_{h/2}^E, \qquad \Phi_h^M = \Phi_{h/2}^E \circ \Phi_{h/2}^I$$

(see Fig. 8.1). This shows $\Phi_h^T = \chi_h^{-1}\Phi_h^M\chi_h$ with $\chi_h = \Phi_{h/2}^E$, implying that the trapezoidal and midpoint rules are mutually conjugate. The change of coordinates, which transforms the numerical solution of one method to that of the other, is $\mathcal{O}(h)$-close to the identity.

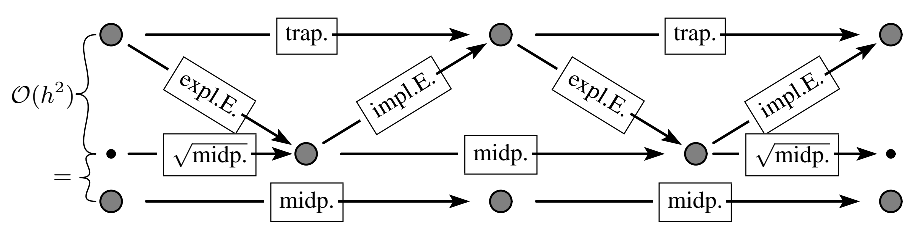


Fig. 8.1. Conjugacy of the trapezoidal rule and the implicit midpoint rule

In fact, we can do even better. If we let $\Phi_{h/2}$ be the square root of Φ_h^M (i.e., $\Phi_{h/2} \circ \Phi_{h/2} = \Phi_h^M$, see Lemma V.3.2), then we have (Fig. 8.1)

$$\Phi_h^T = (\Phi_{h/2}^E)^{-1} \circ \Phi_h^M \circ \Phi_{h/2}^E = (\Phi_{h/2}^E)^{-1} \circ \Phi_{h/2} \circ \Phi_{h/2} \circ \Phi_{h/2} \circ \Phi_{h/2}^{-1} \circ \Phi_{h/2}^E$$

so that the trapezoidal and the midpoint rules are conjugate via $\chi_h = \Phi_{h/2}^{-1} \circ \Phi_{h/2}^E$. Since $\Phi_{h/2}$ and $\Phi_{h/2}^E$ are both consistent with the same differential equation, the transformation χ_h is $\mathcal{O}(h^2)$-close to the identity. This shows that for every numerical solution of the trapezoidal rule there exists a numerical solution of the midpoint rule which remains $\mathcal{O}(h^2)$-close as long as it stays in a compact set. A single trajectory of the non-symplectic trapezoidal rule therefore behaves very much the same as a trajectory of the symplectic implicit midpoint rule.

A Study via B-Series. An investigation of Runge–Kutta methods, conjugate to a symplectic method, leads us to the following weaker requirement: we say that a numerical method Φ_h is *conjugate to a symplectic method* Ψ_h *up to order* r, if there exists a transformation $\chi_h(y) = y + \mathcal{O}(h)$ such that

$$\Phi_h(h) = \big(\chi_h^{-1} \circ \Psi_h \circ \chi_h\big)(y) + \mathcal{O}(h^{r+1}). \tag{8.2}$$

This implies that the error of such a method behaves as the superposition of the error of a symplectic method of order p with that of a non-symplectic method of order r.

In the following we assume that all methods considered as well as the conjugacy mapping χ_h can be represented as B-series

$$\Phi_h(y) = B(a,y), \qquad \Psi_h(y) = B(b,y), \qquad \chi_h(y) = B(c,y). \tag{8.3}$$

Using the composition formula (III.1.38) of B-series, condition (8.2) becomes

$$(ac)(\tau) = (cb)(\tau) \qquad \text{for} \quad |\tau| \le r. \tag{8.4}$$

The following results are taken from the thesis of P. Leone (2000).

Theorem 8.2. *Let $\Phi_h(y) = B(a,y)$ represent a numerical method of order 2.*
a) It is always conjugate to a symplectic method up to order 3.
b) It is conjugate to a symplectic method up to order 4, if and only if

$$a(\bullet,[\bullet,\bullet]) - 2a(\bullet,[[\bullet]]) = 0, \qquad a([\bullet],[\bullet]) - 2a(\bullet,[[\bullet]]) = 0. \tag{8.5}$$

Here, we use the abbreviation $a(u,v) = a(u)\cdot a(v) - a(u\circ v) - a(v\circ u)$.

Proof. The condition (8.4) allows us to express $b(\tau)$ as a function of $a(u)$ for $|u| \le |\tau|$ and of $c(v)$ for $|v| \le |\tau| - 1$ (use the formulas of Example III.1.11). All we have to do is to check the symplecticity conditions $b(u,v) = 0$ for $|u| + |v| \le r$ (see Theorem 7.6).

Since the method Φ_h is of order 2, we obtain $b(\bullet) = 1$ and $b([\bullet]) = 1/2$. We arbitrarily fix $c(\bullet) = 0$, so that the symplecticity condition $b(\bullet,[\bullet]) = 0$ becomes $2c([\bullet]) = a(\bullet,[\bullet])$. Defining $c([\bullet])$ by this relation proves statement (a).

For order 4, the three symplecticity conditions $b(\bullet,[\bullet,\bullet]) = b(\bullet,[[\bullet]]) = b([\bullet],[\bullet]) = 0$ have to be fulfilled. One of them can be satisfied by defining suitably $c([\bullet,\bullet]) + c([[\bullet]])$; the other two conditions are then equivalent to (8.5). □

Theorem 8.3. *Let $\Phi_h(y) = B(a,y)$ represent a numerical method of order 4. It is conjugate to a symplectic method up to order 5, if and only if*

$$a(\bullet,[[\bullet,\bullet]]) - 2a(\bullet,[[[\bullet]]]) = 0, \qquad a(\bullet,[\bullet,\bullet,\bullet]) - 3a(\bullet,[[\bullet],\bullet]) + 3a(\bullet,[[[\bullet]]]) = 0,$$
$$a([\bullet],[\bullet,\bullet]) - a(\bullet,[[\bullet],\bullet]) - 2a([\bullet],[[\bullet]]) + 3a(\bullet,[[[\bullet]]]) = 0.$$

Proof. The idea of the proof is the same as in the preceding theorem. The verification is left as an exercise for the reader. □

Example 8.4. A direct computation shows that for the Lobatto IIIB method with $s = 3$ we have $a([\bullet],[\bullet,\bullet]) = 1/144$, and $a(u,v) = 0$ for all other pairs with $|u| + |v| = 5$. Theorem 8.3 therefore proves that this method is not conjugate to a symplectic method up to order 5.

For the Lobatto IIIA method with $s = 3$ we obtain $a([\bullet],[\bullet,\bullet]) = -1/144$, $a([\bullet],[[\bullet]]) = -1/288$, and $a(u,v) = 0$ for the remaining pairs with $|u| + |v| = 5$. This time the conditions of Theorem 8.3 are fulfilled, so that the Lobatto IIIA method with $s = 3$ is conjugate to a symplectic method up to order 5 at least.

VI.8.2 Near Conservation of Quadratic First Integrals

We have already met in Sect. VI.4.1 a close relationship between symplecticity and the conservation of quadratic first integrals. The aim of this section is to show a similar connection between conjugate symplecticity and the near conservation of quadratic first integrals. This has first been observed and proved by Chartier, Faou & Murua (2005) using the algebra of rooted trees.

Let $Q(y) = y^T Cy$ (with symmetric matrix C) be a quadratic first integral of $\dot y = f(y)$, and assume that $\Phi_h(y)$ is conjugate to a method $\Psi_h(y)$ that exactly conserves quadratic first integrals (e.g., symplectic Runge–Kutta methods). This means that $y_{n+1} = \Phi_h(y_n)$ satisfies

$$\chi_h(y_{n+1})^T C\chi_h(y_{n+1}) = \chi_h(y_n)^T C\chi_h(y_n),$$

and the expression $\widetilde Q(y) = \chi_h(y)^T C\chi_h(y)$ is exactly conserved by the numerical solution of $\Phi_h(y)$. If $\chi_h(y) = B(c,y)$ is a B-series, this is of the form

$$\widetilde Q(y) = \sum_{\tau,\vartheta\in T\cup\{\emptyset\}} h^{|\tau|+|\vartheta|}\beta(\tau,\vartheta)F(\tau)(y)^T C\, F(\vartheta)(y), \tag{8.6}$$

where $F(\emptyset)(y) = y$ and $|\emptyset| = 0$ for the empty tree, and $\beta(\emptyset,\emptyset) = 1$. We have the following criterion for conjugate symplecticity, where all formulas have to be interpreted in the sense of formal series.

Theorem 8.5. *Assume that a one-step method $\Phi_h(y) = B(a,y)$ leaves (8.6) invariant for all problems $\dot y = f(y)$ having $Q(y) = y^T Cy$ as first integral.*

Then, it is conjugate to a symplectic integrator $\Psi_h(z)$, i.e., there exists a transformation $z = \chi_h(y) = B(c,y)$ such that $\Psi_h(z) = \chi_h\circ\Phi_h\circ\chi_h^{-1}(z)$, or equivalently, $\Psi_h(z) = B(c^{-1}ac,z)$ is symplectic.

Proof. The idea is to search for a B-series $B(c,y)$ such that the expression (8.6) becomes

$$\widetilde Q(y) = B(c,y)^T C\, B(c,y).$$

The mapping $z = \chi_h(y) = B(c,y)$ then provides a change of variables such that the original first integral $Q(z) = z^T Cz$ is invariant in the new variables. By Theorem 7.6 this then implies that Ψ_h is symplectic.

By Lemma 8.6 below, the expression (8.6) can be written as

$$\widetilde Q(y) = y^T C\Big(y + \sum_{\theta\in T} h^{|\theta|}\eta(\theta)F(\theta)(y)\Big), \tag{8.7}$$

where $\eta(\theta) = 0$ for $|\theta| < r$, if the perturbation in (8.6) is of size $\mathcal{O}(h^r)$. Using the same lemma once more, we obtain

$$\begin{aligned} B(c,y)^T C\, B(c,y) \;&=\; y^T C\Big(y + 2\sum_{\theta\in T}\frac{h^{|\theta|}}{\sigma(\theta)}\,c(\theta)F(\theta)(y)\Big) \\ &+\; y^T C\Big(\sum_{\theta\in T}\Big(\frac{h^{|\theta|}}{\sigma(\theta)}\sum_{\tau,\vartheta\in T}\frac{\sigma(\theta)\kappa_{\tau,\vartheta}(\theta)}{\sigma(\tau)\sigma(\vartheta)}\,c(\tau)c(\vartheta)F(\theta)(y)\Big). \end{aligned} \tag{8.8}$$

A comparison of the coefficients in (8.7) and (8.8) uniquely defines $c(\theta)$ in a recursive manner. We have $c(\theta) = 0$ for $|\theta| < r$, so that the transformation $z = B(c, y)$ is $\mathcal{O}(h^r)$ close to the identity. □

The previous proof is based on the following result.

Lemma 8.6. *Let $Q(y) = y^T C y$ (with symmetric matrix C) be a first integral of $\dot{y} = f(y)$. Then, for every pair of trees $\tau, \vartheta \in T$, we have*

$$F(\tau)(y)^T C\, F(\vartheta)(y) = y^T C \Big(\sum_{\theta \in T} \kappa_{\tau,\vartheta}(\theta) F(\theta)(y) \Big).$$

This sum is finite and only over trees satisfying $|\theta| = |\tau| + |\vartheta|$.

Proof. By definition of a first integral we have $y^T C\, f(y) = 0$ for all y. Differentiation with respect to y gives

$$f(y)^T C\, k + y^T C\, f'(y)k = 0 \qquad \text{for all } k. \tag{8.9}$$

Putting $k = F(\vartheta)(y)$, this proves the statement for $\tau = \bullet$.

Differentiating once more yields

$$(f'(y)\ell)^T C\, k + \ell^T C\, f'(y)k + y^T C\, f''(y)(k, \ell) = 0.$$

Putting $\ell = f(y)$ and using (8.9), we get the statement for $\tau = [\bullet]$. With $\ell = F(\tau_1)(y)$ we obtain the statement for $\tau = [\tau_1]$ provided that it is already proved for τ_1. We need a further differentiation to get a similar statement for $\tau = [\tau_1, \tau_2]$, etc. The proof concludes by induction on the order of τ. □

Partitioned Methods. This criterion for conjugate symplecticity can be extended to partitioned P-series methods. For partitioned problems

$$\dot{p} = f_1(p, q), \qquad \dot{q} = f_2(p, q) \tag{8.10}$$

we consider first integrals of the form $L(p, q) = p^T E\, q$, where E is an arbitrary constant matrix. If $\Phi_h(p, q)$ is conjugate to a method that exactly conserves $L(p, q)$, then it will conserve a modified first integral of the form

$$\widetilde{L}(p, q) = \sum_{\tau \in TP_p \cup \{\emptyset_p\}, \vartheta \in TP_q \cup \{\emptyset_q\}} h^{|\tau| + |\vartheta|} \beta(\tau, \vartheta) F(\tau)(p, q)^T E\, F(\vartheta)(p, q), \tag{8.11}$$

where $\beta(\emptyset_p, \emptyset_q) = 1$, $F(\emptyset_p)(p, q) = p$, $F(\emptyset_q)(p, q) = q$. We first extend Lemma 8.6 to the new situation.

Lemma 8.7. *Let $L(p, q) = p^T E\, q$ be a first integral of (8.10). Then, for every pair of trees $\tau \in TP_p, \vartheta \in TP_q$, we have*

$$\begin{aligned} F(\tau)(p, q)^T E\, F(\vartheta)(p, q) \;&=\; p^T E \Big(\sum_{\theta \in TP_q} \kappa_{\tau,\vartheta}(\theta) F(\theta)(p, q) \Big) \\ &+\; \Big(\sum_{\theta \in TP_p} \kappa_{\tau,\vartheta}(\theta) F(\theta)(p, q) \Big)^T E\, q. \end{aligned} \tag{8.12}$$

These sums are finite and only over trees satisfying $|\theta| = |\tau| + |\vartheta|$.

Proof. Since $L(p,q) = p^T E\, q$ is a first integral of the differential equation, we have $f_1(p,q)^T E\, q + p^T E\, f_2(p,q) = 0$ for all p and q. As in the proof of Lemma 8.6 the statement follows from differentiation of this relation. □

Theorem 8.8. *Assume that a partitioned one-step method $\Phi_h(p,q) = P(a,(p,q))$ leaves (8.11) invariant for all problems (8.10) having $L(p,q) = p^T E\, q$ as first integral.*

Then it is conjugate to a symplectic integrator $\Psi_h(u,v)$, i.e., there is a transformation $(u,v) = \chi_h(p,q) = P(c,(p,q))$ such that $\Psi_h(u,v) = \chi_h \circ \Phi_h \circ \chi_h^{-1}(u,v)$, or equivalently, $\Psi_h(u,v) = P(c^{-1}ac,(u,v))$ is symplectic.

Proof. We search for a P-series $P(c,(p,q)) = \big(P_p(c,(p,q)), P_q(c,(p,q))\big)^T$ such that the expression (8.11) can be written as

$$\widetilde{L}(p,q) = P_p(c,(p,q))^T E\, P_q(c,(p,q)).$$

As in the proof of Theorem 8.5 the mapping $(u,v) = \chi_h(p,q) = P(c,(p,q))$ then provides the searched change of variables.

Using Lemma 8.7 the expression (8.11) becomes

$$\widetilde{L}(p,q) = p^T E\Big(q + \sum_{\theta \in TP_q} h^{|\theta|} \eta(\theta) F(\theta)(p,q)\Big) + \Big(\sum_{\theta \in TP_p} h^{|\theta|} \eta(\theta) F(\theta)(p,q)\Big)^T E\, q.$$

Also $P_p(c,(p,q))^T E\, P_q(c,(p,q))$ can be written in such a form, and a comparison of the coefficients yields the coefficients $c(\tau)$ of the P-series $P(c,(p,q))$ in a recursive manner. We again have that $P(c,(p,q))$ is $\mathcal{O}(h^r)$ close to the identity, if the perturbation in (8.11) is of size $\mathcal{O}(h^r)$. □

The statement of Theorem 8.8 remains true in the class of second order differential equations $\ddot{q} = f_1(q)$, i.e., $\dot{p} = f_1(p)$, $\dot{q} = p$.

VI.9 Volume Preservation

The flow φ_t of a Hamiltonian system preserves volume in phase space: for every bounded open set $\Omega \subset \mathbb{R}^{2d}$ and for every t for which $\varphi_t(y)$ exists for all $y \in \Omega$,

$$\mathrm{vol}(\varphi_t(\Omega)) = \mathrm{vol}(\Omega)\ ,$$

where $\mathrm{vol}(\Omega) = \int_\Omega dy$. This identity is often referred to as *Liouville's theorem*. It is a consequence of the transformation formula for integrals and the fact that

$$\det \frac{\partial \varphi_t(y)}{\partial y} = 1 \qquad \text{for all } t \text{ and } y, \tag{9.1}$$

which follows directly from the symplecticity and $\varphi_0 = \mathrm{id}$. The same argument shows that every symplectic transformation, and in particular every symplectic integrator applied to a Hamiltonian system, preserves volume in phase space.

More generally than for Hamiltonian systems, volume is preserved by the flow of differential equations with a divergence-free vector field:

Lemma 9.1. *The flow of a differential equation* $\dot{y} = f(y)$ *in* $\mathbb{R}^n$ *is volume-preserving if and only if* $\operatorname{div} f(y) = 0$ *for all* y.

Proof. The derivative $Y(t) = \frac{\partial \varphi_t}{\partial y}(y_0)$ is the solution of the variational equation

$$\dot{Y} = A(t)Y\,, \qquad Y(0) = I\,,$$

with the Jacobian matrix $A(t) = f'(y(t))$ at $y(t) = \varphi_t(y_0)$. From the proof of Lemma IV.3.1 we obtain the *Abel–Liouville–Jacobi–Ostrogradskii identity*

$$\frac{d}{dt}\,\det Y = \text{trace}\, A(t) \cdot \det Y. \tag{9.2}$$

Note that here trace $A(t) = \operatorname{div} f(y(t))$. Hence, $\det Y(t) = 1$ for all t if and only if $\operatorname{div} f(y(t)) = 0$ for all t. Since this is valid for all choices of initial values y_0, the result follows. □

Example 9.2 (ABC Flow). This flow, named after the three independent authors Arnold, Beltrami and Childress, is given by the equations

$$\begin{aligned} \dot{x} &= A \sin z + C \cos y \\ \dot{y} &= B \sin x + A \cos z \\ \dot{z} &= C \sin y + B \cos x \end{aligned} \tag{9.3}$$

and has all diagonal elements of f' identically zero. It is therefore volume preserving. In Arnold (1966, p. 347) it appeared in a footnote as an example of a flow with rot$f$ parallel to f, thus violating Arnold's condition for the existence of invariant tori (Arnold 1966, p. 346). It was therefore expected to possess interesting chaotic properties and has since then been the object of many investigations showing their non-integrability (see e.g., Ziglin (1996)). We illustrate in Fig. 9.1 the action of this flow by transforming, in a volume preserving manner, a ball in $\mathbb{R}^3$. We see that, very soon, the set is strongly squeezed in one direction and dilated in two others. The solutions thus depend in a very sensitive way on the initial values.

Volume-Preserving Numerical Integrators. The question arises as to whether volume-preserving integrators can be constructed for every differential equation with volume-preserving flow. Already for linear problems, Lemma IV.3.2 shows that no standard method can be volume-preserving for dimension $n \geq 3$. Nevertheless, positive answers were found by Qin & Zhu (1993), Shang (1994a,1994b), Feng & Shang (1995) and Quispel (1995). In the following we present the approach of Feng & Shang (1995). The key is the following result which generalizes and reinterprets a construction of H. Weyl (1940) for $n = 3$.

Theorem 9.3 (Feng & Shang 1995). *Every divergence-free vector field* $f : \mathbb{R}^n \to \mathbb{R}^n$ *can be written as the sum of* $n-1$ *vector fields*

$$f = f_{1,2} + f_{2,3} + \ldots + f_{n-1,n}$$

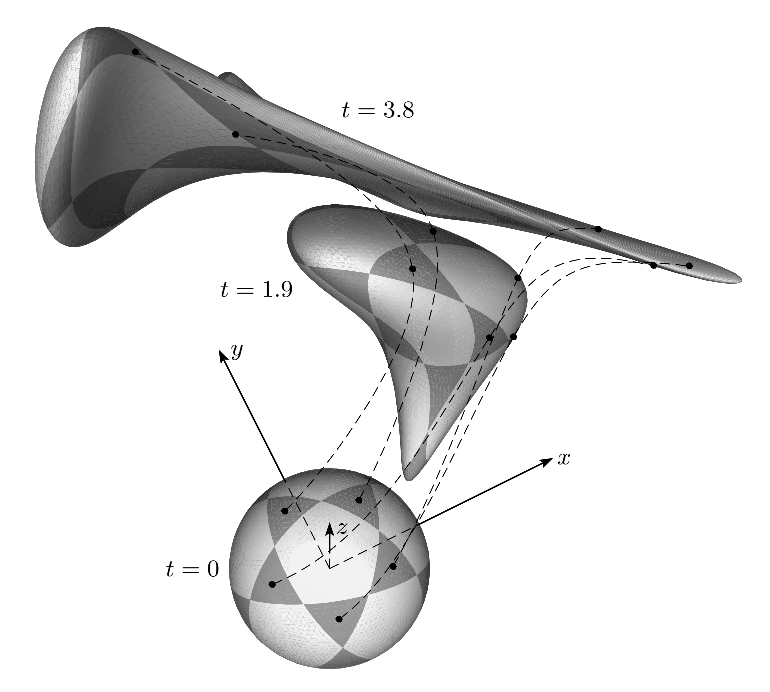


Fig. 9.1. Volume preserving deformation of the ball of radius 1, centred at the origin, by the ABC flow; $A = 1/2$, $B = C = 1$

where each $f_{k,k+1}$ *is Hamiltonian in the variables* (y_k, y_{k+1})*: there exist functions* $H_{k,k+1} : \mathbb{R}^n \to \mathbb{R}$ *such that*

$$f_{k,k+1} = (0, \ldots, 0, -\frac{\partial H_{k,k+1}}{\partial y_{k+1}}, \frac{\partial H_{k,k+1}}{\partial y_k}, 0, \ldots, 0)^T .$$

Proof. In terms of the components of $f = (f_1, \ldots, f_n)^T$, the functions $H_{k,k+1}$ must satisfy the equations

$$f_1 = -\frac{\partial H_{1,2}}{\partial y_2}, \quad f_2 = \frac{\partial H_{1,2}}{\partial y_1} - \frac{\partial H_{2,3}}{\partial y_3}, \ldots,$$

$$f_{n-1} = \frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - \frac{\partial H_{n-1,n}}{\partial y_n}, \quad f_n = \frac{\partial H_{n-1,n}}{\partial y_{n-1}} .$$

We thus set

$$H_{1,2} = -\int_0^{y_2} f_1 \, dy_2$$

and for $k = 2, \ldots, n-2$

$$H_{k,k+1} = \int_0^{y_{k+1}} \Big(\frac{\partial H_{k-1,k}}{\partial y_{k-1}} - f_k \Big) \, dy_{k+1} .$$

It remains to construct $H_{n-1,n}$ from the last two equations. We see by induction that for $k \le n-2$,

$$\frac{\partial^2 H_{k,k+1}}{\partial y_k \partial y_{k+1}} = -\Big(\frac{\partial f_1}{\partial y_1} + \ldots + \frac{\partial f_k}{\partial y_k}\Big) ,$$

and hence the integrability condition for $H_{n-1,n}$,

$$\frac{\partial}{\partial y_{n-1}}\Big(\frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - f_{n-1}\Big) = \frac{\partial f_n}{\partial y_n} ,$$

reduces to the condition $\mathrm{div} f = 0$, which is satisfied by assumption. $H_{n-1,n}$ can thus be constructed as

$$H_{n-1,n} = \int_0^{y_n} \Big(\frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - f_{n-1}\Big)\, dy_n + \int_0^{y_{n-1}} f_n|_{y_n=0}\, dy_{n-1} ,$$

which completes the proof. □

The above construction also shows that

$$f_{k,k+1} = (0, \ldots, 0, f_k + g_k, -g_{k+1}, 0, \ldots, 0)$$

with

$$g_{k+1} = \int_0^{y_{k+1}} \Big(\frac{\partial f_1}{\partial y_1} + \ldots + \frac{\partial f_k}{\partial y_k}\Big)\, dy_{k+1}$$

for $1 \le k \le n-2$, and $g_1 = 0$ and $g_n = -f_n$.

With the decomposition of Lemma 9.3 at hand, a volume-preserving algorithm is obtained by applying a splitting method with symplectic substeps. For example, as proposed by Feng & Shang (1995), a second-order volume-preserving method is obtained by Strang splitting with symplectic Euler substeps:

$$\varphi_h \approx \Phi_h = \Phi_{h/2}^{[1,2]*} \circ \ldots \circ \Phi_{h/2}^{[n-1,n]*} \circ \Phi_{h/2}^{[n-1,n]} \circ \ldots \circ \Phi_{h/2}^{[1,2]}$$

where $\Phi_{h/2}^{[k,k+1]}$ is a symplectic Euler step of length $h/2$ applied to the system with right-hand side $f_{k,k+1}$, and $*$ denotes the adjoint method. In this method, one step $\widehat{y} = \Phi_h(y)$ is computed component-wise, in a Gauss-Seidel-like manner, as

$$\begin{aligned}
\overline{y}_1 &= y_1 + \frac{h}{2} f_1(\overline{y}_1, y_2, \ldots, y_n) \\
\overline{y}_k &= y_k + \frac{h}{2} f_k(\overline{y}_1, \ldots, \overline{y}_k, y_{k+1}, \ldots, y_n) + \frac{h}{2} g_k|_{y_k}^{\overline{y}_k} \quad \text{for } k = 2, \ldots, n-1 \\
\overline{y}_n &= y_n + \frac{h}{2} f_n(\overline{y}_1, \ldots, \overline{y}_{n-1}, y_n)
\end{aligned} \tag{9.4}$$

with $g_k|_{y_k}^{\overline{y}_k} = g_k(\overline{y}_1, \ldots, \overline{y}_k, y_{k+1}, \ldots, y_n) - g_k(\overline{y}_1, \ldots, \overline{y}_{k-1}, y_k, \ldots, y_n)$, and

$$\begin{aligned}
\widehat{y}_n &= \overline{y}_n + \frac{h}{2} f_n(\overline{y}_1, \ldots, \widehat{y}_n) \\
\widehat{y}_k &= \overline{y}_k + \frac{h}{2} f_k(\overline{y}_1, \ldots, \overline{y}_k, \widehat{y}_{k+1} \ldots, \widehat{y}_n) - \frac{h}{2} \overline{g}_k|_{\overline{y}_k}^{\widehat{y}_k} \quad \text{for } k = n-1, \ldots, 2 \\
\widehat{y}_1 &= \overline{y}_1 + \frac{h}{2} f_1(\overline{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n)
\end{aligned} \tag{9.5}$$

with $\overline{g}_k|_{\overline{y}_k}^{\widehat{y}_k} = g_k(\overline{y}_1, \ldots, \overline{y}_{k-1}, \widehat{y}_k, \ldots, \widehat{y}_n) - g_k(\overline{y}_1, \ldots, \overline{y}_k, \widehat{y}_{k+1}, \ldots, \widehat{y}_n)$. The method is one-dimensionally implicit in general, but becomes explicit in the particular case where $\partial f_k / \partial y_k = 0$ for all k.

Separable Partitioned Systems. For problems of the form

$$\dot{y} = f(z), \qquad \dot{z} = g(y) \tag{9.6}$$

with $y \in \mathbb{R}^m$, $z \in \mathbb{R}^n$, the scheme (9.4) becomes the symplectic Euler method, (9.5) its adjoint, and its composition the Lobatto IIIA - IIIB extension of the Störmer–Verlet method. Since symplectic explicit partitioned Runge–Kutta methods are compositions of symplectic Euler steps (Theorem VI.4.7), this observation proves that such methods are volume-preserving for systems (9.6). This fact was obtained by Suris (1996) by a direct calculation, without interpreting the methods as composition methods. The question arises as to whether more symplectic partitioned Runge–Kutta methods are volume-preserving for systems (9.6).

Theorem 9.4. *Every symplectic Runge–Kutta method with at most two stages is volume-preserving for systems (9.6) of arbitrary dimension.*

Proof. (a) The idea is to consider the Hamiltonian system with

$$H(u, v, y, z) = u^T f(z) + v^T g(y),$$

where (u, v) are the conjugate variables to (y, z). This system is of the form

$$\begin{aligned}
\dot{y} &= f(z) & \qquad \dot{u} &= -g'(y)^T v \\
\dot{z} &= g(y) & \qquad \dot{v} &= -f'(z)^T u.
\end{aligned} \tag{9.7}$$

Applying the Runge–Kutta method to this augmented system does not change the numerical solution for (y, z). For symplectic methods the matrix

$$\Big(\frac{\partial(y_1, z_1, u_1, v_1)}{\partial(y_0, z_0, u_0, v_0)} \Big) = M = \begin{pmatrix} R & 0 \\ S & T \end{pmatrix} \tag{9.8}$$

satisfies $M^T J M = J$ which implies $RT^T = I$. Below we shall show that $\det T = \det R$. This yields $\det R = 1$ which implies that the method is volume preserving.

(b) *One-stage methods*. The only symplectic one-stage method is the implicit midpoint rule for which R and T are computed as

$$\Big(I - \frac{h}{2}\,E_1\Big)R \;=\; I + \frac{h}{2}\,E_1 \tag{9.9}$$

$$\Big(I + \frac{h}{2}\,E_1^T\Big)T \;=\; I - \frac{h}{2}\,E_1^T, \tag{9.10}$$

where E_1 is the Jacobian of the system (9.6) evaluated at the internal stage value. Since

$$E_1 = \begin{pmatrix} 0 & f'(z_{1/2}) \\ g'(y_{1/2}) & 0 \end{pmatrix},$$

a similarity transformation with the matrix $D = \mathrm{diag}(I, -I)$ takes E_1 to $-E_1$. Hence, the transformed matrix satisfies

$$\Big(I - \frac{h}{2}\,E_1^T\Big)(D^{-1}TD) \;=\; I + \frac{h}{2}\,E_1^T.$$

A comparison with (9.9) and the use of $\det X^T = \det X$ proves $\det R = \det T$ for the midpoint rule.

(c) *Two-stage methods*. Applying a two-stage implicit Runge–Kutta method to (9.7) yields

$$\begin{pmatrix} I - ha_{11}E_1 & -ha_{12}E_2 \\ -ha_{21}E_1 & I - ha_{22}E_2 \end{pmatrix} \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} \;=\; \begin{pmatrix} I \\ I \end{pmatrix},$$

where R_i is the derivative of the (y, z) components of the ith stage with respect to (y_0, z_0), and E_i is the Jacobian of the system (9.6) evaluated at the ith internal stage value. From the solution of this system the derivative R of (9.8) is obtained as

$$R \;=\; I + (b_1E_1, b_2E_2) \begin{pmatrix} I - ha_{11}E_1 & -ha_{12}E_2 \\ -ha_{21}E_1 & I - ha_{22}E_2 \end{pmatrix}^{-1} \begin{pmatrix} I \\ I \end{pmatrix}.$$

With the determinant identity

$$det(U)\,\det(X - WU^{-1}V) \;=\; \det \begin{pmatrix} U & V \\ W & X \end{pmatrix} \;=\; det(X)\,\det(U - VX^{-1}W),$$

which is seen by Gaussian elimination, this yields

$$\det R \;=\; \frac{\det\big(I \otimes I - h((A - 1\!\!1 b^T) \otimes I)\,E\big)}{\det\big(I \otimes I - h(A \otimes I)\,E\big)},$$

where A and b collect the Runge–Kutta coefficients, and $E =$ blockdiag (E_1, E_2). For $D^{-1}TD$ we get the same formula with E replaced by E^T. If A is an arbitrary 2×2 matrix, it follows from block Gaussian elimination that

$$\det\big(I \otimes I - h(A \otimes I)\,E\big) \;=\; \det\big(I \otimes I - h(A \otimes I)\,E^T\big), \tag{9.11}$$

which then proves $\det R = \det T$. Notice that the identity (9.11) is no longer true in general if A is of dimension larger than two. □

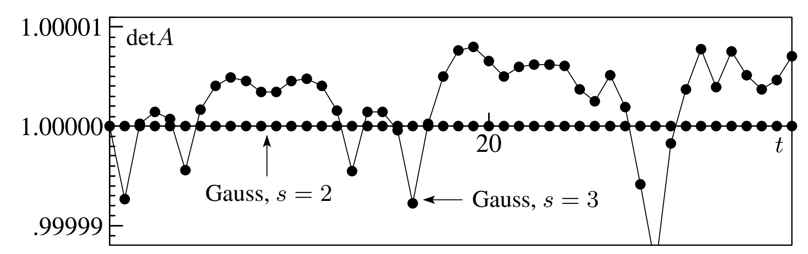


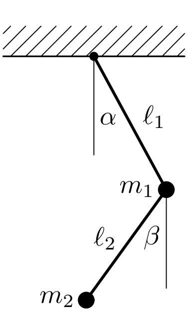
Fig. 9.2. Volume preservation of Gauss methods applied to (9.12) with $h = 0.8$

We are curious to see whether Theorem 9.4 remains valid for symplectic Runge–Kutta methods with more than two stages. For this we apply the Gauss methods with $s = 2$ and $s = 3$ to the problem

$$\dot{x} = \sin z, \qquad \dot{y} = \cos z, \qquad \dot{z} = \sin y + \cos x \tag{9.12}$$

with initial value $(0, 0, 0)$. We show in Fig. 9.2 the determinant of the derivative of the numerical flow as a function of time. Only the two-stage method is volume-preserving for this problem which is in agreement with Theorem 9.4.

VI.10 Exercises



1. Let α and β be the generalized coordinates of the double pendulum, whose kinetic and potential energies are

$$\begin{aligned} T &= \frac{m_1}{2}(\dot{x}_1^2 + \dot{y}_1^2) + \frac{m_2}{2}(\dot{x}_2^2 + \dot{y}_2^2) \\ U &= m_1 g y_1 + m_2 g y_2. \end{aligned}$$

   Determine the generalized momenta of the corresponding Hamiltonian system.

2. A non-autonomous Hamiltonian system is given by a time-dependent Hamiltonian function $H(p, q, t)$ and the differential equations

$$\dot{p} = -H_q(p, q, t), \qquad \dot{q} = H_p(p, q, t).$$

   Verify that these equations together with $\dot{e} = -H_t(p, q, t)$ and $\dot{t} = 1$ are the canonical equations for the extended Hamiltonian $\widetilde{H}(\widetilde{p}, \widetilde{q}) = H(p, q, t) + e$ with $\widetilde{p} = (p, e)$ and $\widetilde{q} = (q, t)$.
3. Prove that a linear transformation $A : \mathbb{R}^2 \to \mathbb{R}^2$ is symplectic, if and only if $\det A = 1$.
4. Consider the transformation $(r, \varphi) \mapsto (p, q)$, defined by

$$p = \psi(r) \cos \varphi, \qquad q = \psi(r) \sin \varphi.$$

   For which function $\psi(r)$ is it a symplectic transformation?

5. Prove that the definition (2.4) of $\Omega(M)$ does not depend on the parametrization φ, i.e., the parametrization $\psi = \varphi \circ \alpha$, where α is a diffeomorphism between suitable domains of $\mathbb{R}^2$, leads to the same result.
6. On the set $U = \{(p, q)\,;\, p^2 + q^2 > 0\}$ consider the differential equation

$$\begin{pmatrix} \dot p \\ \dot q \end{pmatrix} = \frac{1}{p^2 + q^2} \begin{pmatrix} p \\ q \end{pmatrix}. \tag{10.1}$$

Prove that
a) its flow is symplectic everywhere on U;
b) on every simply-connected subset of U the vector field (10.1) is Hamiltonian (with $H(p, q) = -\text{Im}\,\log(p + iq) + Const$);
c) it is not possible to find a differentiable function $H : U \to \mathbb{R}$ such that (10.1) is equal to $J^{-1}\nabla H(p, q)$ for all $(p, q) \in U$.
Remark. The vector field (10.1) is locally (but not globally) Hamiltonian.
7. (Burnton & Scherer 1998). Prove that all members of the one-parameter family of Nyström methods of order $2s$, constructed in Exercise III.9, are symplectic and symmetric.
8. Prove that the statement of Lemma 4.1 remains true for methods that are formally defined by a B-series, $\Phi_h(y) = B(a, y)$.
9. Compute the generating function $S^1(P, q, h)$ of a symplectic Nyström method applied to $\ddot q = U(q)$.
10. Find the Hamilton–Jacobi equation (cf. Theorem 5.7) for the generating function $S^2(p, Q)$ of Lemma 5.3.
11. (*Jacobi's method for exact integration*). Suppose we have a solution $S(q, Q, t, \alpha)$ of the Hamilton–Jacobi equation (5.16), depending on d parameters $\alpha_1, \ldots, \alpha_d$ such that the matrix $\big(\frac{\partial^2 S}{\partial \alpha_i \partial Q_j}\big)$ is invertible. Since this matrix is the Jacobian of the system

$$\frac{\partial S}{\partial \alpha_i} = 0 \qquad i = 1, \ldots, d, \tag{10.2}$$

this system determines a solution path $Q_1, \ldots, Q_q$ which is locally unique. In possession of an additional parameter (and, including the partial derivatives with respect to t, an additional row and column in the Hessian matrix condition), we can also determine $Q_j(t)$ as function of t. Apply this method to the Kepler problem (I.2.2) in polar coordinates, where, with the generalized momenta $p_r = \dot r$, $p_\varphi = r^2 \dot\varphi$, the Hamiltonian becomes

$$H = \frac{1}{2}\Big(p_r^2 + \frac{p_\varphi^2}{r^2}\Big) - \frac{M}{r}$$

and the Hamilton–Jacobi differential equation (5.16) is

$$S_t + \frac{1}{2}\big(S_r\big)^2 + \frac{1}{2r^2}\big(S_\varphi\big)^2 - \frac{M}{r} = 0.$$

Solve this equation by the ansatz $S(t, r, \varphi) = \theta_1(t) + \theta_2(r) + \theta_3(\varphi)$ (separation of variables).

Result. One obtains

$$S = \int \sqrt{2\alpha_1 r^2 + 2Mr - \alpha_2^2}\,\frac{dr}{r} + \alpha_2 \varphi - \alpha_1 t.$$

Putting, e.g., $\partial S/\partial \alpha_2 = 0$, we obtain $\varphi = \arcsin \frac{Mr-\alpha_2^2}{\sqrt{M^2+2\alpha_1\alpha_2^2}\, r}$ by evaluating an elementary integral. This, when resolved for r, leads to the elliptic movement of Kepler (Sect. I.2.2). This method turned out to be most effective for the exact integration of difficult problems. With the same ideas, just more complicated in the computations, Jacobi solves in "lectures" 24 through 30 of (Jacobi 1842) the Kepler motion in $\mathbb{R}^3$, the geodesics of ellipsoids (his greatest triumph), the motion with two centres of gravity, and proves a theorem of Abel.

12. (*Chan's Lobatto IIIS methods.*) Show that there exists a one-parameter family of symplectic, symmetric (and A-stable) Runge–Kutta methods of order $2s-2$ based on Lobatto quadrature (Chan 1990). A special case of these methods can be obtained by taking the arithmetic mean of the Lobatto IIIA and Lobatto IIIB method coefficients (Sun 2000).
Hint. Use the W-transformation (see Hairer & Wanner (1996), p. 77) by putting $X_{s,s-1} = -X_{s-1,s}$ an arbitrary constant.
13. For a Hamiltonian system with associated Lagrangian $L(q,\dot q) = \frac{1}{2}\dot q^T M \dot q - U(q)$, show that every first integral $I(p,q) = p^T a(q)$ resulting from Noether's Theorem has a linear $a(q) = Aq + c$ with skew-symmetric MA.
Hint. (a) It is sufficient to consider the case $M = I$.
(b) Show that $a'(q)$ is skew-symmetric.
(c) Let $a_{ij}(q) = \frac{\partial a_i}{\partial q_j}(q)$. Using the symmetry of the Hessian of each component $a_i(q)$, show that $a_{ij}(q)$ does not depend on q_i, q_j, and is at most linear in the remaining components q_k. With the skew-symmetry of $a'(q)$, conclude that $a'(q) = Const.$
14. Consider the unconstrained *optimal control problem*

$$\begin{aligned} &C\big(q(T)\big) \;\rightarrow\; \min \\ &\dot q(t) = f\big(q(t),u(t)\big), \quad q(0) = q_0 \end{aligned} \tag{10.3}$$

on the interval $[0,T]$, where the control function is assumed to be continuous. Prove that first-order necessary optimality conditions can be written as

$$\begin{aligned} \dot q(t) &= \nabla_p H\big(p(t),q(t),u(t)\big), & q(0) &= q_0 \\ \dot p(t) &= -\nabla_q H\big(p(t),q(t),u(t)\big), & p(T) &= \nabla_q C\big(q(T)\big) \\ 0 &= \nabla_u H\big(p(t),q(t),u(t)\big), \end{aligned} \tag{10.4}$$

where the Hamiltonian is given by

$$H(p,q,u) = p^T f(q,u)$$

(we assume that the Hessian $\nabla_u^2 H(p,q,u)$ is invertible, so that the third relation of (10.4) defines u as a function of (p,q)).

Hint. Consider a slightly perturbed control function $u(t) + \varepsilon\delta u(t)$, and let $q(t)+\varepsilon\delta q(t)+\mathcal{O}(\varepsilon^2)$ be the corresponding solution of the differential equation in (10.3). With the function $p(t)$ of (10.4) we then have

$$C'\big(q(T)\big)\,\delta q(T) \;=\; \int_0^T \frac{d}{dt}\Big(p(t)^T\delta q(t)\Big)dt = \int_0^T p(t)^T f_u\big(\ldots\big)\delta u(t)dt.$$

The algebraic relation of (10.4) then follows from the fundamental lemma of variational calculus.

15. A Runge–Kutta discretization of the problem (10.3) is

$$\begin{aligned} C(q_N) \;&\rightarrow\; \min \\ q_{n+1} \;&=\; q_n + h\textstyle\sum_{i=1}^s b_i f(Q_{ni},U_{ni}) \\ Q_{ni} \;&=\; q_n + h\textstyle\sum_{j=1}^s a_{ij} f(Q_{nj},U_{nj}) \end{aligned} \tag{10.5}$$

with $n = 0,\ldots,N-1$ and $h = T/N$. We assume $b_i \neq 0$ for all i. Introducing suitable Lagrange multipliers for the constrained minimization problem (10.5), prove that there exist p_n, P_{ni} such that the optimal solution of (10.5) satisfies (Hager 2000)

$$\begin{aligned} q_{n+1} &= q_n + h\textstyle\sum_{i=1}^s b_i\nabla_p H(P_{ni},Q_{ni},U_{ni}) \\ Q_{ni} &= q_n + h\textstyle\sum_{j=1}^s a_{ij}\nabla_p H(P_{nj},Q_{nj},U_{nj}) \\ p_{n+1} &= p_n - h\textstyle\sum_{i=1}^s \widehat b_i\nabla_q H(P_{ni},Q_{ni},U_{ni}) \\ P_{ni} &= p_n - h\textstyle\sum_{j=1}^s \widehat a_{ij}\nabla_q H(P_{nj},Q_{nj},U_{nj}) \\ 0 &= \nabla_u H(P_{ni},Q_{ni},U_{ni}) \end{aligned} \tag{10.6}$$

with $p_N = \nabla_q C(q_N)$ and given initial value q_0, where the coefficients $\widehat b_i$ and $\widehat a_{ij}$ are determined by

$$\widehat b_i = b_i, \qquad b_i\widehat a_{ij} + \widehat b_j a_{ji} = b_i\widehat b_j. \tag{10.7}$$

Consequently, (10.6) can be considered as a symplectic discretization of (10.4); see Bonnans & Laurent-Varin (2006).

16. (Hager 2000). For an explicit s-stage Runge–Kutta method of order $p = s$ and $b_i \neq 0$, consider the partitioned Runge–Kutta method with additional coefficients $\widehat b_i$ and $\widehat a_{ij}$ defined by (10.7). Prove the following:
a) For $p = s = 3$, the partitioned method is of order 3 if and only if $c_3 = 1$.
b) For $p = s = 4$, the partitioned method is of order 4 without any restriction.

Chapter VII. Non-Canonical Hamiltonian Systems

We discuss theoretical properties and the structure-preserving numerical treatment of Hamiltonian systems on manifolds and of the closely related class of Poisson systems. We present numerical integrators for problems from classical and quantum mechanics.

VII.1 Constrained Mechanical Systems

Constrained mechanical systems form an important class of differential equations on manifolds. Their numerical treatment has been extensively investigated in the context of *differential-algebraic equations* and is documented in monographs like that of Brenan, Campbell & Petzold (1996), Eich-Soellner & Führer (1998), Hairer, Lubich & Roche (1989), and Chap. VII of Hairer & Wanner (1996). We concentrate here on the symmetry and/or symplecticity of such numerical integrators.

VII.1.1 Introduction and Examples

Consider a mechanical system described by position coordinates $q_1, \ldots, q_d$, and suppose that the motion is constrained to satisfy $g(q) = 0$ where $g : \mathbb{R}^d \to \mathbb{R}^m$ with $m < d$. Let $T(q, \dot q) = \frac{1}{2}\dot q^T M(q) \dot q$ be the kinetic energy of the system and $U(q)$ its potential energy, and put

$$L(q, \dot q) = T(q, \dot q) - U(q) - g(q)^T \lambda, \tag{1.1}$$

where $\lambda = (\lambda_1, \ldots, \lambda_m)^T$ consists of Lagrange multipliers. The Euler–Lagrange equation of the variational problem for $\int_0^t L(q, \dot q)\, dt$ is then given by

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot q}\Big) - \frac{\partial L}{\partial q} \;=\; 0.$$

Written as a first order differential equation we get

$$\begin{aligned} \dot q &= v \\ M(q)\, \dot v &= f(q, v) - G(q)^T \lambda \\ 0 &= g(q), \end{aligned} \tag{1.2}$$

where $f(q, v) = -\frac{\partial}{\partial q}\big(M(q)v\big)v + \nabla_q T(q, v) - \nabla_q U(q)$ and $G(q) = \frac{\partial g}{\partial q}(q)$.

Example 1.1 (Spherical Pendulum). We denote by q_1, q_2, q_3 the Cartesian coordinates of a point with mass m that is connected with a massless rod of length ℓ to the origin. The kinetic and potential energies are $T = \frac{m}{2}(\dot q_1^2 + \dot q_2^2 + \dot q_3^2)$ and $U = mgq_3$, respectively, and the constraint is the fixed length of the rod. We thus get the system

$$\begin{aligned} \dot q_1 &= v_1 & m\,\dot v_1 &= -2q_1\lambda \\ \dot q_2 &= v_2 & m\,\dot v_2 &= -2q_2\lambda \\ \dot q_3 &= v_3 & m\,\dot v_3 &= -mg - 2q_3\lambda \\ 0 &= q_1^2 + q_2^2 + q_3^2 - \ell^2. \end{aligned} \tag{1.3}$$

The physical meaning of λ is the tension in the rod which maintains the constant distance of the mass point from the origin.

Existence and Uniqueness of the Solution. A standard approach for studying the existence of solutions of differential-algebraic equations is to differentiate the constraints until an ordinary differential equation is obtained. Differentiating the constraint in (1.2) twice with respect to time yields

$$0 = G(q)\,v \qquad \text{and} \qquad 0 = g''(q)(v,v) + G(q)\,\dot v. \tag{1.4}$$

The equation for $\dot v$ in (1.2) together with the second relation of (1.4) constitute a linear system for $\dot v$ and λ,

$$\begin{pmatrix} M(q) & G(q)^T \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} \dot v \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,v) \\ -g''(q)(v,v) \end{pmatrix}. \tag{1.5}$$

Throughout this chapter we require the matrix appearing in (1.5) to be invertible for q close to the solution we are looking for. This then allows us to express $\dot v$ and λ as functions of (q, v). Notice that the matrix in (1.5) is invertible when $G(q)$ has full rank and $M(q)$ is invertible on $\ker G(q) = \{h \mid G(q)h = 0\}$.

We are now able to discuss the existence of a solution of (1.2). First of all, observe that the initial values q_0, v_0, λ_0 cannot be arbitrarily chosen. They have to satisfy the first relation of (1.4) and $\lambda_0 = \lambda(q_0, v_0)$, where $\lambda(q, v)$ is obtained from (1.5). In the case that q_0, v_0, λ_0 satisfy these conditions, we call them *consistent initial values*. Furthermore, every solution of (1.2) has to satisfy

$$\dot q = v, \qquad \dot v = \dot v(q, v), \tag{1.6}$$

where $\dot v(q, v)$ is the function obtained from (1.5). It is known from standard theory of ordinary differential equations that (1.6) has locally a unique solution. This solution $\big(q(t), v(t)\big)$ together with $\lambda(t) := \lambda\big(q(t), v(t)\big)$ satisfies (1.5) by construction, and hence also the two differential equations of (1.2). Integrating the second relation of (1.4) twice and using the fact that the integration constants vanish for consistent initial values, proves also the remaining relation $0 = g(q)$ for this solution.

Formulation as a Differential Equation on a Manifold. We denote by

$$\mathcal{Q} = \{q\,;\,g(q)=0\} \tag{1.7}$$

the *configuration manifold*, on which the positions q are constrained to lie. The tangent space at $q \in \mathcal{Q}$ is $T_q\mathcal{Q} = \{v\,;\,G(q)v=0\}$. The equations (1.6) define thus a differential equation on the manifold

$$T\mathcal{Q} = \big\{(q,v)\,;\,q\in\mathcal{Q},\ v\in T_q\mathcal{Q}\big\} = \big\{(q,v)\,;\,g(q)=0,\,G(q)v=0\big\}, \tag{1.8}$$

the *tangent bundle* of $\mathcal{Q}$. Indeed, we have just shown that for initial values $(q_0, v_0) \in T\mathcal{Q}$ (i.e., consistent initial values) the problems (1.6) and (1.2) are equivalent, so that the solutions of (1.6) stay on $T\mathcal{Q}$.

Reversibility. The system (1.2) and the corresponding differential equation (1.6) are reversible with respect to the involution $\rho(q,v) = (q,-v)$, if $f(q,-v) = f(q,v)$. This follows at once from Example V.1.3, because the solution $\dot v(q,v)$ of (1.5) satisfies $\dot v(q,-v) = \dot v(q,v)$

For the numerical solution of differential-algebraic equations "index reduction" is a very popular technique. This means that instead of directly treating the problem (1.2) one numerically solves the differential equation (1.6) on the manifold $\mathcal{M}$. Projection methods (Sect. IV.4) as well as methods based on local coordinates (Sect. IV.5) are much in use. If one is interested in a correct simulation of the reversible structure of the problem, the symmetric methods of Sect. V.4 can be applied. Here we do not repeat these approaches for this particular situation, instead we concentrate on the symplectic integration of constrained systems.

VII.1.2 Hamiltonian Formulation

In Sect. VI.1 we have seen that, for unconstrained mechanical systems, the equations of motion become more structured if we use the momentum coordinates $p = \frac{\partial L}{\partial \dot q} = M(q)\dot q$ in place of the velocity coordinates $v = \dot q$. Let us do the same for the constrained system (1.2). As in the proof of Theorem VI.1.3 we obtain the equivalent system

$$\begin{aligned} \dot q &= H_p(p,q) \\ \dot p &= -H_q(p,q) - G(q)^T\lambda \\ 0 &= g(q), \end{aligned} \tag{1.9}$$

where

$$H(p,q) = \frac{1}{2}\,p^T M(q)^{-1} p + U(q) \tag{1.10}$$

is the total energy of the system; H_p and H_q denote the column vectors of partial derivatives. Differentiating the constraint in (1.9) twice with respect to time, we get

$$0 \;=\; G(q)H_p(p,q), \tag{1.11}$$

$$0 \;=\; \frac{\partial}{\partial q}\Big(G(q)H_p(p,q)\Big)H_p(p,q) - G(q)H_{pp}(p,q)\Big(H_q(p,q) + G(q)^T\lambda\Big), \tag{1.12}$$

and assuming the matrix

$$G(q)H_{pp}(p,q)G(q)^T \qquad \text{is invertible,} \tag{1.13}$$

equation (1.12) permits us to express λ in terms of (p,q).

Formulation as a Differential Equation on a Manifold. Inserting the so-obtained function $\lambda(p,q)$ into (1.9) gives a differential equation for (p,q) on the manifold

$$\mathcal{M} = \big\{(p,q)\,;\, g(q)=0,\ G(q)H_p(p,q)=0\big\}. \tag{1.14}$$

As we will now see, this manifold has a differential-geometric interpretation as the cotangent bundle of the configuration manifold $\mathcal{Q} = \{q\,;\, g(q) = 0\}$. The Lagrangian for a fixed $q \in \mathcal{Q}$ is a function on the tangent space $T_q\mathcal{Q}$, i.e., $L(q,\cdot) : T_q\mathcal{Q} \to \mathbb{R}$. Its (Fréchet) derivative evaluated at $\dot q \in T_q\mathcal{Q}$ is therefore a linear mapping $d_{\dot q}L(q,\dot q) : T_q\mathcal{Q} \to \mathbb{R}$, or in other terms, $d_{\dot q}L(q,\dot q)$ is in the cotangent space $T_q^*\mathcal{Q}$. Since the duality is such that $\langle d_{\dot q}L(q,\dot q), v\rangle = \frac{\partial L}{\partial \dot q}(q,\dot q)v$ for $v \in T_q\mathcal{Q}$, condition (1.13) ensures that the Legendre transform $\dot q \mapsto p = d_{\dot q}L(q,\dot q)$ is an invertible transformation between $T_q\mathcal{Q}$ and $T_q^*\mathcal{Q}$. We can therefore consider $T_q^*\mathcal{Q}$ as a subspace of $\mathbb{R}^d$ if every $p \in T_q^*\mathcal{Q}$ is identified with $\frac{\partial L}{\partial \dot q}(q,\dot q)^T = M(q)\dot q \in \mathbb{R}^d$ for the unique $\dot q \in T_q\mathcal{Q}$ for which $p = d_{\dot q}L(q,\dot q)$ holds. With this identification,

$$T_q^*\mathcal{Q} = \{M(q)\dot q\,;\, \dot q \in T_q\mathcal{Q}\},$$

and the duality is given by $\langle p, v\rangle = p^Tv$ for $p \in T_q^*\mathcal{Q}$ and $v \in T_q\mathcal{Q}$. We thus have $p = M(q)\dot q \in T_q^*\mathcal{Q}$ if and only if $\dot q = M(q)^{-1}p = H_p(p,q) \in T_q\mathcal{Q}$. Since the tangent space at $q \in \mathcal{Q}$ is $T_q\mathcal{Q} = \{\dot q\,;\, G(q)\dot q = 0\}$, we obtain that

$$p \in T_q^*\mathcal{Q} \qquad \text{if and only if} \qquad G(q)H_p(p,q) = 0.$$

Denoting by $T^*\mathcal{Q} = \{(p,q)\,;\, q \in \mathcal{Q},\ p \in T_q^*\mathcal{Q}\}$ the *cotangent bundle* of $\mathcal{Q}$, we thus see that the constraint manifold $\mathcal{M}$ of (1.14) equals

$$\mathcal{M} = T^*\mathcal{Q}. \tag{1.15}$$

The constrained Hamiltonian system (1.9) with Hamiltonian (1.10) can thus be viewed as a differential equation on the cotangent bundle $T^*\mathcal{Q}$ of the configuration manifold $\mathcal{Q}$.

In the following we consider the system (1.9)–(1.12) with (1.13) where $H(p,q)$ is an arbitrary smooth function. The constraint manifold is then still given by (1.14). The existence and uniqueness of the solution of (1.9) can be discussed as before.

Reversibility. It is readily checked that the system (1.9) is reversible if $H(-p,q) = H(p,q)$. This is always satisfied for a Hamiltonian (1.10).

Preservation of the Hamiltonian. Differentiation of $H\big(p(t),q(t)\big)$ with respect to time yields

$$-H_p^TH_q - H_p^TG^T\lambda + H_q^TH_p$$

with all expressions evaluated at $\big(p(t), q(t)\big)$. The first and the last terms cancel, and the central term vanishes because $GH_p = 0$ on the solution manifold. Consequently, the Hamiltonian $H(p,q)$ is constant along solutions of (1.9).

Symplecticity of the Flow. Since the flow of the system (1.9) is a transformation on $\mathcal{M}$, its derivative is a mapping between the corresponding tangent spaces. In agreement with Definition VI.2.2 we call a map $\varphi : \mathcal{M} \to \mathcal{M}$ symplectic if, for every $x = (p,q) \in \mathcal{M}$,

$$\xi_1^T \varphi'(x)^T J\, \varphi'(x) \xi_2 = \xi_1^T J \xi_2 \qquad \text{for all} \quad \xi_1, \xi_2 \in T_x\mathcal{M}. \tag{1.16}$$

If φ is actually defined and continuously differentiable in an open subset of $\mathbb{R}^{2d}$ that contains $\mathcal{M}$, then $\varphi'(x)$ in the above formula is just the usual Jacobian matrix. Otherwise, some care is necessary in the interpretation of (1.16): φ' is the tangent map given by the directional derivative $\varphi'(x)\xi := (d/d\tau)|_{\tau=0}\ \varphi(\gamma(\tau))$ for $\xi \in T_x\mathcal{M}$, where γ is a path on $\mathcal{M}$ with $\gamma(0) = x$, $\dot\gamma(0) = \xi$. The expression $\xi_1^T \varphi'(x)^T$ in (1.16) should then be interpreted as $(\varphi'(x)\xi_1)^T$.

Theorem 1.2. *Let $H(p,q)$ and $g(q)$ be twice continuously differentiable. The flow $\varphi_t : \mathcal{M} \to \mathcal{M}$ of the system (1.9) is then a symplectic transformation on $\mathcal{M}$, i.e., it satisfies (1.16).*

Proof. We let $x = (p,q)$, so that the system (1.9) becomes $\dot x = J^{-1}\big(\nabla H(x) + \sum_i \lambda_i(x) \nabla g_i(x)\big)$, where $\lambda_i(x)$ and $g_i(x)$ are the components of $\lambda(x)$ and $g(x)$, and $\lambda(x)$ is the function obtained from (1.12). The variational equation of this system, satisfied by the directional derivative $\Psi = \varphi_t'(x_0)\xi$, with $x_0 = (p_0, q_0)$, reads

$$\dot\Psi = J^{-1}\Big(\nabla^2 H(x) + \sum_{i=1}^m \lambda_i(x) \nabla^2 g_i(x) + \sum_{i=1}^m \nabla g_i(x) \nabla\lambda_i(x)^T\Big)\Psi.$$

A direct computation, analogous to that in the proof of Theorem VI.2.4, yields for $\xi_1, \xi_2 \in T_x\mathcal{M}$

$$\begin{aligned}\frac{d}{dt}\Big(\xi_1^T \varphi_t'(x_0)^T J\, \varphi_t'(x_0)\xi_2\Big) \;=\; \ldots \;=\; & \sum_{i=1}^m \xi_1^T \varphi_t'(x_0)^T \nabla g_i(x) \nabla\lambda_i(x)^T \varphi_t'(x_0)\xi_2 \\ -\; & \sum_{i=1}^m \xi_1^T \varphi_t'(x_0)^T \nabla\lambda_i(x) \nabla g_i(x)^T \varphi_t'(x_0)\xi_2.\end{aligned} \tag{1.17}$$

Since $g_i\big(\varphi_t(x_0)\big) = 0$ for $x_0 \in \mathcal{M}$, we have $\nabla g_i(x)^T \varphi_t'(x_0)\xi_2 = 0$ and the same for ξ_1, so that the expression in (1.17) vanishes. This proves the symplecticity of the flow on $\mathcal{M}$. □

Differentiating the constraint in (1.9) twice and solving for the Lagrange multiplier from (1.12) (this procedure is known as "index reduction" of the differential-algebraic system) yields the differential equation

$$\dot q = H_p(p,q), \qquad \dot p = -H_q(p,q) - G(q)^T \lambda(p,q), \tag{1.18}$$

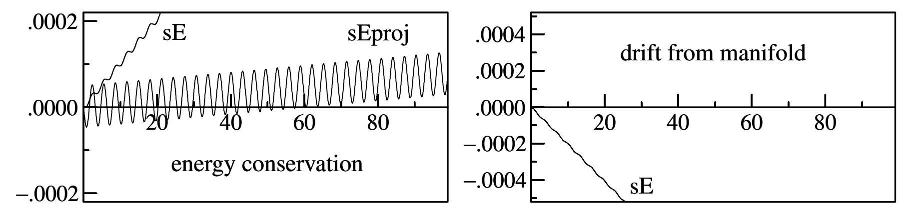


Fig. 1.1. Numerical solution of the symplectic Euler method applied to (1.18) with $H(p,q) = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) + q_3$, $g(q) = q_1^2 + q_2^2 + q_3^2 - 1$ (spherical pendulum); initial value $q_0 = (0, \sin(0.1), -\cos(0.1))$, $p_0 = (0.06, 0, 0)$, step size $h = 0.003$ for method "sE" (without projection) and $h = 0.03$ for method "sEproj" (with projection)

where $\lambda(p,q)$ is obtained from (1.12). If we solve this system with the symplectic Euler method (implicit in p, explicit in q), the qualitative behaviour of the numerical solution is not correct. As was observed by Leimkuhler & Reich (1994), there is a linear error growth in the Hamiltonian and also a drift from the manifold $\mathcal{M}$ (method "sE" in Fig. 1.1). The explanation for this behaviour is the fact that (1.18) is no longer a Hamiltonian system. If we combine the symplectic Euler applied to (1.18) with an orthogonal projection onto $\mathcal{M}$ (method "sEproj"), the result improves considerably but the linear error growth in the Hamiltonian is not eliminated. This numerical experiment illustrates that "index reduction" is not compatible with symplectic integration.

VII.1.3 A Symplectic First Order Method

We extend the symplectic Euler method to Hamiltonian systems with constraints. We integrate the p-variable by the implicit and the q-variable by the explicit Euler method. This gives

$$\begin{aligned} \widehat{p}_{n+1} &= p_n - h\big(H_q(\widehat{p}_{n+1}, q_n) + G(q_n)^T \lambda_{n+1}\big) \\ q_{n+1} &= q_n + h\, H_p(\widehat{p}_{n+1}, q_n) \\ 0 &= g(q_{n+1}). \end{aligned} \tag{1.19}$$

The numerical approximation $(\widehat{p}_{n+1}, q_{n+1})$ satisfies the constraint $g(q) = 0$, but not $G(q)H_p(p,q) = 0$. To get an approximation $(p_{n+1}, q_{n+1}) \in \mathcal{M}$, we append the projection

$$\begin{aligned} p_{n+1} &= \widehat{p}_{n+1} - h\, G(q_{n+1})^T \mu_{n+1} \\ 0 &= G(q_{n+1}) H_p(p_{n+1}, q_{n+1}). \end{aligned} \tag{1.20}$$

Let us discuss some basic properties of this method.

Existence and Uniqueness of the Numerical Solution. Inserting the definition of q_{n+1} from the second line of (1.19) into $0 = g(q_{n+1})$ gives a nonlinear system for $\widehat{p}_{n+1}$ and $h\lambda_{n+1}$. Due to the factor h in front of $H_p(\widehat{p}_{n+1}, q_n)$, the implicit function theorem cannot be directly applied to prove existence and uniqueness of the numerical solution. We therefore write this equation as

$$0 \,=\, g(q_{n+1}) \,=\, g(q_n) + \int_0^1 G\big(q_n + \tau(q_{n+1} - q_n)\big)(q_{n+1} - q_n)\, d\tau.$$

We now use $g(q_n) \,=\, 0$, insert the definition of q_{n+1} from the second line of (1.19) and divide by h. Together with the first line of (1.19) this yields the system $F\big(\widehat{p}_{n+1}, h\lambda_{n+1}, h\big) = 0$ with

$$F\big(p,\nu,h\big) = \begin{pmatrix} p - p_n + hH_q(p,q_n) + G(q_n)^T\nu \\ \displaystyle\int_0^1 G\big(q_n + \tau h H_p(p,q_n)\big)\, H_p(p,q_n)\, d\tau \end{pmatrix}.$$

Since $(p_n, q_n) \in \mathcal{M}$ with $\mathcal{M}$ from (1.14), we have $F(p_n, 0, 0) = 0$. Furthermore,

$$\frac{\partial F}{\partial(p,\nu)}\big(p_n,0,0\big) = \begin{pmatrix} I & G(q_n)^T \\ G(q_n)H_{pp}(p_n,q_n) & 0 \end{pmatrix},$$

and this matrix is invertible by (1.13). Consequently, an application of the implicit function theorem proves that the numerical solution $(\widehat{p}_{n+1}, h\lambda_{n+1})$ (and hence also q_{n+1}) exists and is locally unique for sufficiently small h.

The projection step (1.20) constitutes a nonlinear system for p_{n+1} and $h\mu_{n+1}$, to which the implicit function theorem can be directly applied.

Convergence of Order 1. The above use of the implicit function theorem yields the rough estimates

$$\widehat{p}_{n+1} = p_n + \mathcal{O}(h), \quad h\lambda_{n+1} = \mathcal{O}(h), \quad h\mu_{n+1} = \mathcal{O}(h),$$

which, together with the equations (1.19) and (1.20), give

$$q_{n+1} = q(t_{n+1}) + \mathcal{O}(h^2), \quad p_{n+1} = p(t_{n+1}) - G\big(q(t_{n+1})\big)^T\nu + \mathcal{O}(h^2),$$

where $\big(p(t), q(t)\big)$ is the solution of (1.9) passing through $(p_n, q_n) \in \mathcal{M}$ at $t = t_n$. Inserting these relations into the second equation of (1.20) we get

$$0 = G\big(q(t)\big)H_p\big(p(t),q(t)\big) + G\big(q(t)\big)H_{pp}\big(p(t),q(t)\big)G\big(q(t)\big)^T\nu + \mathcal{O}(h^2)$$

at $t \,=\, t_{n+1}$. Since $G\big(q(t)\big)H_p\big(p(t),q(t)\big) \,=\, 0$, it follows from (1.13) that $\nu \,=\, O(h^2)$. The local error is therefore of size $\mathcal{O}(h^2)$.

The convergence proof now follows standard arguments, because the method is a mapping $\Phi_h : \mathcal{M} \to \mathcal{M}$ on the solution manifold. We consider the solutions $\big(p_n(t), q_n(t)\big)$ of (1.9) passing through the numerical values $(p_n, q_n) \in \mathcal{M}$ at $t = t_n$, we estimate the difference of two successive solutions in terms of the local error at t_n, and we sum up the propagated errors (see Fig. 3.2 of Sect. II.3 in Hairer, Nørsett & Wanner (1993)). This proves that the global error satisfies $p_n - p(t_n) = \mathcal{O}(h)$ and $q_n - q(t_n) = \mathcal{O}(h)$ as long as $t_n = nh \le Const$.

Symplecticity. We first study the mapping $(p_n, q_n) \mapsto (\widehat{p}_{n+1}, q_{n+1})$ defined by (1.19), and we consider λ_{n+1} as a function $\lambda(p_n, q_n)$. Differentiation with respect to (p_n, q_n) yields

$$\begin{pmatrix} I + hH_{qp}^T & 0 \\ -hH_{pp} & I \end{pmatrix} \left(\frac{\partial(\widehat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right) = \begin{pmatrix} I - hG^T\lambda_p & S - hG^T\lambda_q \\ 0 & I + hH_{qp} \end{pmatrix}, \quad (1.21)$$

where $S = -hH_{qq} - h\lambda^T g_{qq}$ is a symmetric matrix, the expressions H_{qp}, H_{pp}, H_{qq}, G are evaluated at $(\widehat{p}_{n+1}, q_n)$, and λ, λ_p, λ_q at (p_n, q_n). A computation, identical to that of the proof of Theorem VI.3.3, yields

$$\left(\frac{\partial(\widehat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right)^T J \left(\frac{\partial(\widehat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right) = \begin{pmatrix} 0 & I - h\lambda_p^T G \\ -I + hG^T\lambda_p & h(G^T\lambda_q - \lambda_q^T G) \end{pmatrix}.$$

We multiply this relation from the left by $\xi_1 \in T_{(p_n,q_n)}\mathcal{M}$ and from the right by $\xi_2 \in T_{(p_n,q_n)}\mathcal{M}$. With the partitioning $\xi = (\xi_p, \xi_q)$ we have $G(q_n)\xi_{q,j} = 0$ for $j = 1, 2$ so that the expression reduces to $\xi_1^T J \xi_2$. This proves the symplecticity condition (1.16) for the mapping $(p_n, q_n) \mapsto (\widehat{p}_{n+1}, q_{n+1})$.

Similarly, the projection step $(\widehat{p}_{n+1}, q_{n+1}) \mapsto (p_{n+1}, q_{n+1})$ of (1.20) gives

$$\frac{\partial(p_{n+1}, q_{n+1})}{\partial(\widehat{p}_{n+1}, q_{n+1})} = \begin{pmatrix} I - hG^T\mu_p & S - hG^T\mu_q \\ 0 & I \end{pmatrix},$$

where μ_{n+1} of (1.20) is considered as a function of $(\widehat{p}_{n+1}, q_{n+1})$, and $S = -h\mu^T g_{qq}$. This is formally the same as (1.21) with $H \equiv 0$. Consequently, the symplecticity condition is also satisfied for this mapping. As a composition of two symplectic transformations, the numerical flow of our first order method is therefore also symplectic.

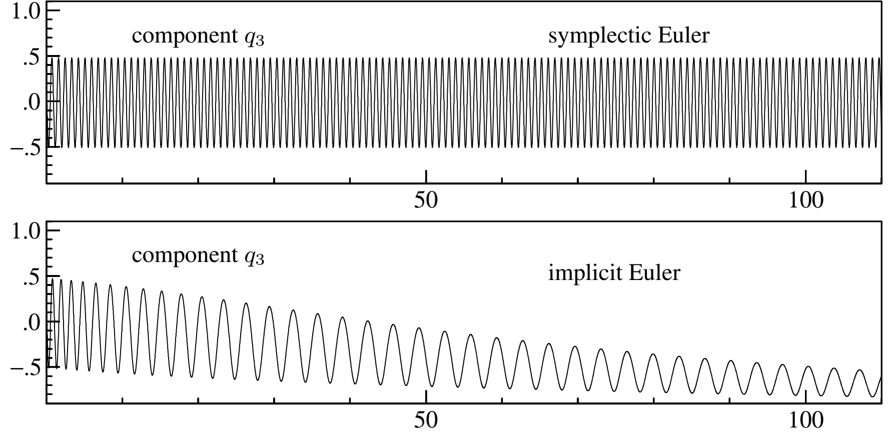


Fig. 1.2. Spherical pendulum problem solved with the symplectic Euler method (1.19)-(1.20) and with the implicit Euler method; initial value $q_0 = (\sin(1.3), 0, \cos(1.3))$, $p_0 = (3\cos(1.3), 6.5, -3\sin(1.3))$, step size $h = 0.01$

Numerical Experiment. Consider the equations (1.3) for the spherical pendulum. For a mass $m = 1$ they coincide with the Hamiltonian formulation. Figure 1.2 (upper picture) shows the numerical solution (vertical coordinate q_3) over many periods obtained by method (1.19)-(1.20). We observe a regular qualitatively correct behaviour. For the implicit Euler method (i.e., the argument q_n is replaced with q_{n+1} in (1.19)) the numerical solution, obtained with the same step size and the same initial values, is less satisfactory. Already after one period the solution deteriorates and the pendulum loses energy.

VII.1.4 SHAKE and RATTLE

The numerical method (1.19)-(1.20) is only of order 1 and it is not symmetric. An algorithm that is of order 2, symmetric and symplectic was originally considered for separable Hamiltonians

$$H(p,q) = \frac{1}{2}\, p^T M^{-1} p + U(q) \tag{1.22}$$

with constant mass matrix M. Notice that in this case we are concerned with a second order differential equation $M\ddot{q} = -U_q(q) - G(q)^T\lambda$ with $g(q) = 0$.

SHAKE. Ryckaert, Ciccotti & Berendsen (1977) propose the method

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= -h^2 M^{-1}\big(U_q(q_n) + G(q_n)^T\lambda_n\big) \\ 0 &= g(q_{n+1}) \end{aligned} \tag{1.23}$$

for computations in molecular dynamics. It is a straightforward extension of the Störmer–Verlet scheme (I.1.15). The p-components, not used in the recursion, are approximated by $p_n = M(q_{n+1} - q_{n-1})/2h$.

RATTLE. The three-term recursion (1.23) may lead to an accumulation of round-off errors, and a reformulation as a one-step method is desirable. Using the same procedure as in (I.1.17) we formally get

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\big(U_q(q_n) + G(q_n)^T\lambda_n\big) \\ q_{n+1} &= q_n + hM^{-1}p_{n+1/2}, \qquad 0 = g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\big(U_q(q_{n+1}) + G(q_{n+1})^T\lambda_{n+1}\big). \end{aligned} \tag{1.24}$$

The difficulty with this formulation is that λ_{n+1} is not yet available at this step (it is computed together with q_{n+2}). As a remedy, Andersen (1983) suggests replacing the last line in (1.24) with a projection step similar to (1.20)

$$\begin{aligned} p_{n+1} &= p_{n+1/2} - \frac{h}{2}\big(U_q(q_{n+1}) + G(q_{n+1})^T\mu_n\big) \\ 0 &= G(q_{n+1})M^{-1}p_{n+1}. \end{aligned} \tag{1.25}$$

This modification, called RATTLE, has the further advantage that the numerical approximation (p_{n+1}, q_{n+1}) lies on the solution manifold $\mathcal{M}$. The symplecticity of this algorithm has been established by Leimkuhler & Skeel (1994).

Extension to General Hamiltonians. As observed independently by Jay (1994) and Reich (1993), the RATTLE algorithm can be extended to general Hamiltonians as follows: for consistent values $(p_n, q_n) \in \mathcal{M}$ define

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\big(H_q(p_{n+1/2}, q_n) + G(q_n)^T\lambda_n\big) \\ q_{n+1} &= q_n + \frac{h}{2}\big(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})\big) \\ 0 &= g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\big(H_q(p_{n+1/2}, q_{n+1}) + G(q_{n+1})^T\mu_n\big) \\ 0 &= G(q_{n+1})H_p(p_{n+1}, q_{n+1}). \end{aligned} \tag{1.26}$$

The first three equations of (1.26) are very similar to (1.19) and the last two equations to (1.20). The existence of (locally) unique solutions $(p_{n+1/2}, q_{n+1}, \lambda_n)$ and (p_{n+1}, μ_n) can therefore be proved in the same way. Notice also that this method gives a numerical solution that stays exactly on the solution manifold $\mathcal{M}$.

Theorem 1.3. *The numerical method (1.26) is symmetric, symplectic, and convergent of order two.*

Proof. Although this theorem is the special case $s = 2$ of Theorem 1.4, we outline its proof. We will see that the convergence result is easier to obtain for $s = 2$ than for the general case.

If we add to (1.26) the consistency conditions $g(q_n) = 0$, $G(q_n)H_p(p_n, q_n) = 0$ of the initial values, the symmetry of the method follows at once by exchanging $h \leftrightarrow -h$, $p_{n+1} \leftrightarrow p_n$, $q_{n+1} \leftrightarrow q_n$, and $\lambda_n \leftrightarrow \mu_n$. The symplecticity can be proved as for (1.19)-(1.20) by computing the derivative of (p_{n+1}, q_{n+1}) with respect to (p_n, q_n), and by verifying the condition (1.16). This does not seem to be simpler than the symplecticity proof of Theorem 1.4.

The implicit function theorem applied to the two subsystems of (1.26) shows

$$p_{n+1/2} = p_n + \mathcal{O}(h), \quad h\lambda = \mathcal{O}(h), \quad p_{n+1} = p_{n+1/2} + \mathcal{O}(h), \quad h\mu = \mathcal{O}(h),$$

and, inserted into (1.26), yields

$$q_{n+1} = q(t_{n+1}) + \mathcal{O}(h^2), \quad p_{n+1} = p(t_{n+1}) - G\big(q(t_{n+1})\big)^T\nu + \mathcal{O}(h^2).$$

Convergence of order one follows therefore in the same way as for method (1.19)-(1.20). Since the order of a symmetric method is always even, this implies convergence of order two. □

An easy way of obtaining high order methods for constrained Hamiltonian systems is by composition (Reich 1996a). Method (1.26) is an ideal candidate as basic integrator for compositions of the form (V.3.2). The resulting integrators are symmetric, symplectic, of high order, and yield a numerical solution that stays on the manifold $\mathcal{M}$.

VII.1.5 The Lobatto IIIA - IIIB Pair

Another possibility for obtaining high order symplectic integrators for constrained Hamiltonian systems is by the use of partitioned Runge–Kutta or discontinuous collocation methods. We consider the system (1.9) and we search for polynomials $u(t)$ of degree s, $w(t)$ of degree $s-1$, and $v(t)$ of degree $s-2$ such that

$$u(t_n) = q_n, \qquad v(t_n) = p_n - hb_1\delta(t_n) \tag{1.27}$$

with the defect

$$\delta(t) = \dot v(t) + H_q\big(v(t),u(t)\big) + G\big(u(t)\big)^T w(t) \tag{1.28}$$

and, using the abbreviation $t_{n,i} = t_n + c_i h$,

$$\begin{aligned} \dot u(t_{n,i}) &= H_p\big(v(t_{n,i}),u(t_{n,i})\big), \quad i=1,\ldots,s \\ \dot v(t_{n,i}) &= -H_q\big(v(t_{n,i}),u(t_{n,i})\big) - G\big(u(t_{n,i})\big)^T w(t_{n,i}), \quad i=2,\ldots,s-1 \\ 0 &= g\big(u(t_{n,i})\big), \quad i=1,\ldots,s. \end{aligned} \tag{1.29}$$

If these polynomials exist, the numerical solution is defined by

$$\begin{aligned} q_{n+1} = u(t_n+h), \qquad p_{n+1} = v(t_n+h) - hb_s\delta(t_n+h) \\ 0 = G(q_{n+1})H_p(p_{n+1},q_{n+1}). \end{aligned} \tag{1.30}$$

Why Discontinuous Collocation Based on Lobatto Quadrature? At a first glance (Theorem VI.4.2) it seems natural to consider collocation methods based on Gaussian quadrature for the entire system. This, however, has the disadvantage that the numerical solution does not satisfy $g(q_{n+1}) = 0$. To achieve this requirement, $t_n + h$ has to be one of the collocation points, i.e., we must have $c_s = 1$. Unfortunately, none of the collocation or discontinuous collocation methods with $c_s = 1$ is symplectic (see Exercise IV.6). We therefore turn our attention to partitioned methods, and we treat only the q-component by a collocation method satisfying $c_s = 1$. To satisfy the s conditions $g\big(u(t_{n,i})\big) = 0$ of (1.29) there are only $s-1$ free parameters $w(t_n), w(t_n + c_2h), \ldots, w(t_n + c_{s-1}h)$ available. A remedy is to choose $c_1 = 0$ so that the first condition $g\big(u(t_n)\big) = 0$ is automatically verified. Encouraged by Theorem VI.4.5 we are thus led to consider the Lobatto nodes in the role of the c_i. The use of the partitioned Lobatto IIIA - IIIB pair for the treatment of constrained Hamiltonian systems has been suggested by Jay (1994, 1996).

Existence and Uniqueness of the Numerical Solution. The polynomial $u(t)$ of degree s is uniquely determined by $u(t_n) = q_n$ and $\dot u(t_{n,i}) =: \dot Q_i$ ($i = 1,\ldots,s$), the polynomial $v(t)$ of degree $s-2$ is uniquely determined by $v(t_{n,i}) =: P_i$ ($i = 1,\ldots,s-1$), and the polynomial $w(t)$ of degree $s-1$ is uniquely determined by $hw(t_{n,i}) =: \Lambda_i$ ($i = 1,\ldots,s$). Notice that the value Λ_s is only involved in (1.30) and not in (1.27)-(1.29). For the nonlinear system (1.27)-(1.29) we therefore consider

$$X = \big(\dot{Q}_1,\ldots,\dot{Q}_s, P_1,\ldots,P_{s-1}, \Lambda_1,\ldots,\Lambda_{s-1}\big)$$

as independent variables, and we write the system as $F(X,h) = 0$. The function F is composed of the s conditions for $\dot{u}(t_{n,i})$, of the definition of $v(t_n)$ (divided by b_1) and the $s-2$ conditions for $\dot{v}(t_{n,i})$ (multiplied by h), and finally of the $s-1$ equations $0 = g\big(u(t_{n,i})\big)$ for $i = 2,\ldots,s$ (divided by h). Observe that $0 = g\big(u(t_n)\big)$ is automatically satisfied by the consistency of (p_n,q_n). We note that $P_s = v(t_n+h)$ and $\dot{P}_i = h\dot{v}(t_{n,i})$ are linear combinations of $P_1,\ldots,P_{s-1}$ with coefficients independent of the step size h.

The function $F(X,h)$ is well-defined for h in a neighbourhood of 0. For the first two blocks this is evident, for the last one it follows from the identity

$$\frac{1}{h}\, g\big(u(t_{n,i})\big) = \int_0^{c_i} G\big(u(t_n+\theta h)\big)\, \dot{u}(t_n+\theta h)\, d\theta$$

using the fact that $\dot{u}(t_n+\theta h)$ is a linear combination of $\dot{Q}_i$ for $i = 1,\ldots,s$. With the values

$$X_0 = \big(H_p(p_n,q_n),\ldots,H_p(p_n,q_n), p_n,\ldots,p_n, 0,\ldots,0\big)$$

we have that $F(X_0,0) = 0$, because the values (p_n,q_n) are assumed to be consistent. In view of an application of the implicit function theorem we compute

$$\frac{\partial F}{\partial X}\big(X_0,0\big) = \begin{pmatrix} I\otimes I & -D\otimes H_{pp} & 0 \\ 0 & B\otimes I & I\otimes G^T \\ A\otimes G & 0 & 0 \end{pmatrix}, \tag{1.31}$$

where H_{pp}, G are evaluated at (p_n,q_n), and A,B,D are matrices of dimension $(s-1)\times s$, $(s-1)\times(s-1)$ and $s\times(s-1)$ respectively that depend only on the Lobatto quadrature and not on the differential equation. For example, the matrix B represents the linear mapping

$$\big(P_1,\ldots,P_{s-1}\big) \mapsto \big(\dot{P}_1 + b_1^{-1}P_1, \dot{P}_2,\ldots,\dot{P}_{s-1}\big).$$

This mapping is invertible, because the values on the right-hand side uniquely determine the polynomial $v(t)$ of degree $s-2$.

Block Gaussian elimination then shows that (1.31) is invertible if and only if the matrix

$$ADB^{-1}\otimes GH_{pp}G^T \qquad \text{is invertible.}$$

Because of (1.13) it remains to show that ADB^{-1} is invertible.

To achieve this without explicitly computing the matrices A,B,D, we apply the method to the problem where p and q are of dimension one, $H(p,q) = p^2/2$, and $g(q) = q$. Assuming $h = 1$ we get

$$\begin{aligned} &u(0) = 0, \qquad v(0) = -b_1\big(\dot{v}(0) + w(0)\big) \\ &\dot{u}(c_i) = v(c_i) \qquad \text{for } i = 1,\ldots,s \\ &\dot{v}(c_i) = -w(c_i) \qquad \text{for } i = 2,\ldots,s-1 \\ &0 = u(c_i) \qquad \text{for } i = 1,\ldots,s, \end{aligned} \tag{1.32}$$

which is equivalent to

$$\begin{pmatrix} I & -D & 0 \\ 0 & B & I \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} (\dot{u}(c_i))_{i=1}^{s} \\ (v(c_i))_{i=1}^{s-1} \\ (w(c_i))_{i=1}^{s-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{1.33}$$

because $H_{pp}(p,q) = 1$ and $G(q) = 1$. Since $u(t)$ is a polynomial of degree s, the last equation of (1.32) implies that $u(t) = C\prod_{j=1}^{s}(t-c_j)$. By the second relation the polynomial $\dot{u}(t) - v(t)$, which is of degree $s-1$, vanishes at s points. Hence, $v(t) \equiv \dot{u}(t)$, which is possible only if $C = 0$, because the degree of $v(t)$ is $s-2$. Consequently, the linear system (1.33) has only the trivial solution, so that the matrix in (1.33) and hence also ADB^{-1} is invertible.

The implicit function theorem applied to $F(X,h) = 0$ shows that the nonlinear system (1.27)-(1.30) possesses a locally unique solution for sufficiently small step sizes h. Using the free parameter $\Lambda_s = hw(t_n + h)$, a further application of the implicit function theorem, this time to the small system (1.30), proves the existence and local uniqueness of p_{n+1}.

Theorem 1.4. *Let $(b_i, c_i)_{i=1}^{s}$ be the weights and nodes of the Lobatto quadrature (c.f. (II.1.17)). The method (1.27)-(1.29)-(1.30) is symmetric, symplectic, and superconvergent of order $2s-2$.*

Proof. Symmetry. To the formulas (1.27)-(1.29)-(1.30) we add the consistency relations $g(q_n) = 0$, $G(q_n)H_p(p_n,q_n) = 0$. Then we exchange $(t_n,p_n,q_n) \leftrightarrow (t_{n+1},p_{n+1},q_{n+1})$ and $h \leftrightarrow -h$. Since $b_1 = b_s$ and $c_{s+1-i} = 1-c_i$ for the Lobatto quadrature, the resulting formulas are equivalent to the original method (see also the proof of Theorem V.2.1).

Symplecticity. We fix $\xi_1, \xi_2 \in T_{(p_n,q_n)}\mathcal{M}$, we put $x_n = (p_n,q_n)^T$, and we consider the bilinear mapping

$$Q\Big(\frac{\partial p_{n+1}}{\partial x_n}, \frac{\partial q_{n+1}}{\partial x_n}\Big) = \xi_1^T \bigg(\Big(\frac{\partial q_{n+1}}{\partial x_n}\Big)^T\Big(\frac{\partial p_{n+1}}{\partial x_n}\Big) - \Big(\frac{\partial p_{n+1}}{\partial x_n}\Big)^T\Big(\frac{\partial q_{n+1}}{\partial x_n}\Big)\bigg)\xi_2.$$

The symplecticity of the transformation $(p_n,q_n) \mapsto (p_{n+1},q_{n+1})$ on the manifold $\mathcal{M}$ is then expressed by the relation

$$Q\Big(\frac{\partial p_{n+1}}{\partial x_n}, \frac{\partial q_{n+1}}{\partial x_n}\Big) = Q\Big(\frac{\partial p_n}{\partial x_n}, \frac{\partial q_n}{\partial x_n}\Big). \tag{1.34}$$

We now follow closely the proof of Theorem IV.2.3. We consider the polynomials $u(t), v(t), w(t)$ of the method (1.27)-(1.29)-(1.30) as functions of t and $x_n = (p_n,q_n)$, and we compute

$$\begin{aligned} Q\Big(\frac{\partial v(t_{n+1})}{\partial x_n}, \frac{\partial u(t_{n+1})}{\partial x_n}\Big) \;-\; & Q\Big(\frac{\partial v(t_n)}{\partial x_n}, \frac{\partial u(t_n)}{\partial x_n}\Big) \\ &= \int_{t_n}^{t_{n+1}} \frac{dQ}{dt}\Big(\frac{\partial v(t)}{\partial x_n}, \frac{\partial u(t)}{\partial x_n}\Big)\, dt. \end{aligned} \tag{1.35}$$

Since $u(t)$ is a polynomial of degree s and $v(t)$ of degree $s-2$, the integrand in (1.35) is a polynomial in t of degree $2s-3$. It is thus integrated without error by the Lobatto quadrature. By definition these polynomials satisfy the differential equation at the interior collocation points. Therefore, it follows from (1.17) that

$$\frac{dQ}{dt}\Big(\frac{\partial v(t_{n,i})}{\partial x_n},\frac{\partial u(t_{n,i})}{\partial x_n}\Big)=0 \qquad \text{for } \; i=2,\dots,s-1,$$

and that

$$\frac{dQ}{dt}\Big(\frac{\partial v(t_{n,i})}{\partial x_n},\frac{\partial u(t_{n,i})}{\partial x_n}\Big)=Q\Big(\frac{\partial \delta(t_{n,i})}{\partial x_n},\frac{\partial u(t_{n,i})}{\partial x_n}\Big) \quad \text{for } \; i=1 \; \text{ and } \; i=s.$$

Applying the Lobatto quadrature to the integral in (1.35) thus yields

$$hb_1\,Q\Big(\frac{\partial \delta(t_n)}{\partial x_n},\frac{\partial u(t_n)}{\partial x_n}\Big)+hb_s\,Q\Big(\frac{\partial \delta(t_{n+1})}{\partial x_n},\frac{\partial u(t_{n+1})}{\partial x_n}\Big),$$

and the symplecticity relation (1.34) follows in the same way as in the proof of Theorem IV.2.3.

Superconvergence. This is the most difficult part of the proof. We remark that superconvergence of Runge–Kutta methods for differential-algebraic systems of index 3 has been conjectured by Hairer, Lubich & Roche (1989), and a first proof has been obtained by Jay (1993) for collocation methods. In his thesis Jay (1994) proves superconvergence for a more general class of methods, including the Lobatto IIIA - IIIB pair, using a "rooted-tree-type" theory. A sketch of that very elaborate proof is published in Jay (1996). Using the idea of discontinuous collocation, the elegant proof for collocation methods can now be extended to cover the Lobatto IIIA - IIIB pair. In the following we explain how the local error can be estimated.

We consider the polynomials $u(t), v(t), w(t)$ defined in (1.27)-(1.29)-(1.30), and we define defects $\mu(t), \delta(t), \theta(t)$ as follows:

$$\begin{aligned} \dot u(t) &= H_p\big(v(t),u(t)\big)+\mu(t)\\ \dot v(t) &= -H_q\big(v(t),u(t)\big)-G\big(u(t)\big)^T w(t)+\delta(t)\\ 0 &= g\big(u(t)\big)+\theta(t). \end{aligned} \tag{1.36}$$

By definition of the method we have

$$\begin{aligned} \mu(t_n+c_ih) &= 0, \qquad i=1,\dots,s\\ \delta(t_n+c_ih) &= 0, \qquad i=2,\dots,s-1\\ \theta(t_n+c_ih) &= 0, \qquad i=1,\dots,s. \end{aligned} \tag{1.37}$$

We let $q(t), p(t), \lambda(t)$ be the exact solution of (1.9) satisfying $q(t_n)=q_n$, $p(t_n)=p_n$, and we consider the differences

$$\Delta u(t)=u(t)-q(t), \quad \Delta v(t)=v(t)-p(t), \quad \Delta w(t)=w(t)-\lambda(t).$$

Subtracting (1.9) from (1.36) we get by linearization that

$$\begin{aligned} \dot{\Delta u} &= a_{11}(t)\Delta u + a_{12}(t)\Delta v + \mu(t) \\ \dot{\Delta v} &= a_{21}(t)\Delta u + a_{22}(t)\Delta v + a_{23}(t)\Delta w + \delta(t), \end{aligned} \tag{1.38}$$

where $a_{12}(t) = H_{pp}\big(p(t), q(t)\big)$, and where the other $a_{ij}(t)$ are given by similar expressions. We have suppressed quadratic and higher order terms to keep the presentation as simple as possible. They do not influence the convergence result. To eliminate Δw in (1.38), we differentiate the algebraic relations in (1.9) and (1.36) twice, and we subtract them. This yields

$$\begin{aligned} 0 &= F\big(t, \mu(t)\big) + b_1(t)\Delta u + b_2(t)\Delta v + B(t)\Delta w \\ &+ G\big(u(t)\big)H_{pp}\big(v(t), u(t)\big)\delta(t) + G\big(u(t)\big)\dot{\mu}(t) + \ddot{\theta}(t), \end{aligned}$$

where $F(t,\mu)$, $B(t)$, $b_1(t)$, $b_2(t)$ are functions depending on $p(t)$, $q(t)$, $\lambda(t)$, $u(t)$, $v(t)$, $w(t)$, and where $F(t,0) = 0$ and $B(t) \approx G(q_n)H_{pp}(p_n, q_n)G(q_n)^T$. Because of our assumption (1.13) we can extract Δw from this relation, and we insert it into (1.38). In this way we get a linear differential equation for $\Delta u, \Delta v$, which can be solved by the "variation of constants" formula. Using $\Delta u(t_n) = 0$ (by (1.27)), the solution $\Delta v(t_n + h)$ is seen to be of the form

$$\begin{aligned} \Delta v(t_n + h) &= R_{22}(t_n + h, t_n)\Delta v(t_n) + \int_{t_n}^{t_n+h} \Big(R_{21}(t_n + h, t)\mu(t) \\ &+ R_{22}(t_n + h, t)\Big(\delta(t) + \widetilde{F}\big(t, \mu(t)\big) + c_1(t)\dot{\mu}(t) \\ &+ C(t)\Big(G\big(u(t)\big)H_{pp}\big(v(t), u(t)\big)\delta(t) + \ddot{\theta}(t)\Big)\Big)\Big)\, dt, \end{aligned} \tag{1.39}$$

where R_{21} and R_{22} are the lower blocks of the resolvent, and $\widetilde{F}, c_1, C$ are functions as before. To prove that the local error of the p-component

$$p_{n+1} - p(t_n + h) = \Delta v(t_n + h) - h b_s \delta(t_n + h) \tag{1.40}$$

is of size $\mathcal{O}(h^{2s-1})$, we first integrate by parts those expressions in (1.39) which contain a derivative. For example,

$$\int_{t_n}^{t_{n+1}} a(t)\dot{\mu}(t)\, dt = a(t)\mu(t)\Big|_{t_n}^{t_{n+1}} - \int_{t_n}^{t_{n+1}} \dot{a}(t)\mu(t)\, dt = \mathcal{O}(h^{2s-1}),$$

because $\mu(t_n) = \mu(t_n + h) = 0$ by (1.37) and an application of the Lobatto quadrature to the integral at the right-hand side gives zero as result with a quadrature error of size $\mathcal{O}(h^{2s-1})$. Similarly, integrating by parts twice yields

$$\begin{aligned} \int_{t_n}^{t_{n+1}} a(t)\ddot{\theta}(t)\, dt &= a(t)\dot{\theta}(t)\Big|_{t_n}^{t_{n+1}} - \dot{a}(t)\theta(t)\Big|_{t_n}^{t_{n+1}} + \int_{t_n}^{t_{n+1}} \ddot{a}(t)\theta(t)\, dt \\ &= a(t_{n+1})\dot{\theta}(t_{n+1}) - a(t_n)\dot{\theta}(t_n) + \mathcal{O}(h^{2s-1}). \end{aligned}$$

To the other integrals in (1.39) we apply the Lobatto quadrature directly. Since $R_{22}(t_{n+1}, t_{n+1})$ is the identity, this gives

$$\begin{aligned} p_{n+1} - p(t_{n+1}) &= R_{22}(t_{n+1}, t_n)\Big(\Delta v(t_n) + hb_1\delta(t_n)\Big) \qquad (1.41) \\ &+ \widetilde{C}(t_{n+1})\Big(hb_s G\big(u(t_{n+1})\big)H_{pp}\big(v(t_{n+1}), u(t_{n+1})\big)\delta(t_{n+1}) + \dot\theta(t_{n+1})\Big) \\ &+ \widetilde{C}(t_n)\Big(hb_1 G\big(u(t_n)\big)H_{pp}\big(v(t_n), u(t_n)\big)\delta(t_n) - \dot\theta(t_n)\Big) + \mathcal{O}(h^{2s-1}), \end{aligned}$$

where $\widetilde{C}(t) = R(t_{n+1}, t)C(t)$. The term $\Delta v(t_n) + hb_1\delta(t_n)$ vanishes by (1.27), and differentiation of the algebraic relation in (1.36) yields

$$0 = G\big(u(t)\big)\Big(H_p\big(v(t), u(t)\big) + \mu(t)\Big) + \dot\theta(t).$$

As a consequence of (1.27), (1.37) and the consistency of the initial values (p_n, q_n), this gives

$$\begin{aligned} \dot\theta(t_n) &= -G(q_n)H_p\big(p_n - hb_1\delta(t_n), q_n\big) \\ &= hb_1 G(q_n)H_{pp}\big(p_n, q_n\big)\delta(t_n) + \mathcal{O}\big(h^2\delta(t_n)^2\big) \\ &= hb_1 G\big(u(t_n)\big)H_{pp}\big(v(t_n), u(t_n)\big)\delta(t_n) + \mathcal{O}\big(h^2\delta(t_n)^2\big). \end{aligned}$$

Using (1.30) we get in the same way

$$\dot\theta(t_{n+1}) = -hb_s G\big(u(t_{n+1})\big)H_{pp}\big(v(t_{n+1}), u(t_{n+1})\big)\delta(t_{n+1}) + \mathcal{O}\big(h^2\delta(t_{n+1})^2\big).$$

These estimates together show that the local error (1.41) is of size $\mathcal{O}(h^{2s-1}) + \mathcal{O}\big(h^2\delta(t)^2\big)$. The defect $\delta(t)$ vanishes at $s-2$ points in the interval $[t_n, t_{n+1}]$, so that $\delta(t) = \mathcal{O}(h^{s-2})$ for $t \in [t_n, t_{n+1}]$ (for a rigorous proof of this statement one has to apply the techniques of the proof of Theorem II.1.5). Therefore we obtain $p_{n+1} - p(t_{n+1}) = \mathcal{O}(h^{2s-2})$, and by the symmetry of the method also $\mathcal{O}(h^{2s-1})$.

In analogy to (1.39), the variation of constants formula yields also an expression for the local error $q_{n+1} - q(t_{n+1}) = \Delta u(t_{n+1})$. One only has to replace R_{21} and R_{22} with the upper blocks R_{11} and R_{12} of the resolvent. Using $R_{12}(t_{n+1}, t_{n+1}) = 0$, we prove in the same way that the local error of the q-component is of size $\mathcal{O}(h^{2s-1})$.

The estimation of the global error is obtained in the same way as for the first order method (1.19)-(1.20). Since the algorithm is a mapping $\Phi_h : \mathcal{M} \to \mathcal{M}$ on the solution manifold, it is not necessary to follow the technically difficult proofs in the context of differential-algebraic equations. Summing up the propagated local errors proves that the global error satisfies $p_n - p(t_n) = \mathcal{O}(h^{2s-2})$ and $q_n - q(t_n) = \mathcal{O}(h^{2s-2})$ as long as $t_n = nh \le Const$. □

VII.1.6 Splitting Methods

When considering splitting methods for constrained mechanical systems, it should be borne in mind that such systems are differential equations on manifolds (see

Sect. VII.1.2). Splitting methods should therefore be based on a decomposition $f(y) = f^{[1]}(y) + f^{[2]}(y)$, where both $f^{[i]}(y)$ are vector fields on the same manifold as $f(y)$. Let us consider here the Hamiltonian system (1.9) with Hamiltonian

$$H(p,q) \;=\; H^{[1]}(p,q) + H^{[2]}(p,q). \tag{1.42}$$

The manifold for this differential equation is

$$\mathcal{M} = \big\{(p,q) \mid g(q) = 0,\; G(q)H_p(p,q) = 0\big\}. \tag{1.43}$$

Notice that (1.9), when H is simply replaced with $H^{[i]}$, is not a good candidate for splitting methods: the existence of a solution is not guaranteed, and if the solution exists it need not stay on the manifold $\mathcal{M}$. The following lemma indicates how splitting methods should be applied.

Lemma 1.5. *Consider a Hamiltonian (1.42), a function $g(q)$ with $G(q) = g'(q)$, and let the manifold $\mathcal{M}$ be given by (1.43). If (1.13) holds and if*

$$G(q)H_p^{[i]}(p,q) = 0 \qquad \text{for all } \; (p,q) \in \mathcal{M}, \tag{1.44}$$

then the system

$$\begin{aligned} \dot q &= H_p^{[i]}(p,q) \\ \dot p &= -H_q^{[i]}(p,q) - G(q)^T\lambda \\ 0 &= G(q)H_p(p,q) \end{aligned} \tag{1.45}$$

defines a differential equation on the manifold $\mathcal{M}$, and its flow is a symplectic transformation on $\mathcal{M}$.

Proof. Differentiation of the algebraic relation in (1.45) with respect to time, and replacing $\dot q$ and $\dot p$ with their differential equations, yields an explicit relation for $\lambda = \lambda(p,q)$ (as a consequence of (1.13)). Hence, a unique solution of (1.45) exists locally if $G(q_0)H_p(p_0,q_0) = 0$. The assumption (1.44) implies $\frac{d}{dt}g\big(q(t)\big) = 0$. This together with the algebraic relation of (1.45) guarantees that for $(p_0,q_0) \in \mathcal{M}$ the solution stays on the manifold $\mathcal{M}$. The symplecticity of the flow is proved as for Theorem 1.2. □

Suppose now that the Hamiltonian $H(p,q)$ of (1.9) can be split as in (1.42), where both $H^{[i]}(p,q)$ satisfy (1.44). We denote by $\varphi_t^{[i]}$ the flow of the system (1.45). If these flows can be computed analytically, the Lie-Trotter splitting $\varphi_h^{[2]} \circ \varphi_h^{[1]}$ and the Strang splitting $\varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ yield first and second order numerical integrators, respectively. Considering more general compositions as in (II.5.6) and using the coefficients proposed in Sect. V.3, methods of high order are obtained. They give numerical approximations lying on the manifold $\mathcal{M}$, and they are symplectic (also symmetric if the splitting is well chosen).

For the important special case where

$$H(p,q) \;=\; T(p,q) + U(q)$$

is the sum of the kinetic and potential energies, both summands satisfy assumption (1.44). This gives a natural splitting that is often used in practice.

Example 1.6 (Spherical Pendulum). We normalize all constants to 1 (cf. Example 1.1) and we consider the problem (1.9) with

$$H(p,q) = \frac{1}{2}\big(p_1^2 + p_2^2 + p_3^2\big) + q_3, \qquad g(q) = \frac{1}{2}\big(q_1^2 + q_2^2 + q_3^2 - 1\big).$$

We split the Hamiltonian as $H^{[1]}(p,q) = \frac{1}{2}\big(p_1^2+p_2^2+p_3^2\big)$ and $H^{[2]}(p,q) = q_3$, and we solve (1.45) with initial values on the manifold

$$\mathcal{M} = \big\{(p,q) \mid q_1^2 + q_2^2 + q_3^2 - 1 = 0,\; p_1q_1 + p_2q_2 + p_3q_3 = 0\big\}.$$

The kinetic energy $H^{[1]}(p,q)$ leads to the system

$$\dot q = p, \qquad \dot p = -q\lambda, \qquad q^Tp = 0,$$

which gives $\lambda = p_0^Tp_0$, so that the flow $\varphi_t^{[1]}$ is just a planar rotation around the origin. The potential energy $H^{[2]}(p,q)$ leads to

$$\dot q = 0, \qquad \dot p = -(0,0,1)^T - q\lambda, \qquad q^Tp = 0.$$

The flow $\varphi_t^{[2]}$ keeps $q(t)$ constant and changes $p(t)$ linearly with time. Splitting methods give simple, explicit and symplectic time integrators for this problem.

VII.2 Poisson Systems

This section is devoted to an interesting generalization of Hamiltonian systems, where J^{-1} in (VI.2.5) is replaced with a nonconstant matrix $B(y)$. Such structures were introduced by Sophus Lie (1888) and are today called *Poisson systems*. They result, in particular, from Hamiltonian systems on manifolds written in non-canonical coordinates. In a first subsection, however, we discuss the Poisson structure of Hamiltonian systems in canonical form.

VII.2.1 Canonical Poisson Structure

> ... quelques remarques sur la plus profonde découverte de M. Poisson, mais qui, je crois, n'a pas été bien comprise ni par Lagrange, ni par les nombreux géomètres qui l'ont citée, ni par son auteur lui-même.
>
> (C.G.J. Jacobi 1840, p. 350)

The derivative of a function $F(p,q)$ along the flow of a Hamiltonian system

$$\dot{p} = -\frac{\partial H}{\partial q}(p,q), \qquad \dot{q} = \frac{\partial H}{\partial p}(p,q), \tag{2.1}$$

is given by (Lie derivative, see (III.5.3))

$$\frac{d}{dt}F\big(p(t),q(t)\big) \;=\; \sum_{i=1}^{d}\Big(\frac{\partial F}{\partial p_i}\,\dot{p}_i + \frac{\partial F}{\partial q_i}\,\dot{q}_i\Big) \;=\; \sum_{i=1}^{d}\Big(\frac{\partial F}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial H}{\partial q_i}\Big). \tag{2.2}$$

This remarkably symmetric structure motivates the following definition.

Definition 2.1. The (canonical) *Poisson bracket* of two smooth functions $F(p,q)$ and $G(p,q)$ is the function

$$\{F,G\} \;=\; \sum_{i=1}^{d}\Big(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i}\Big), \tag{2.3}$$

or in vector notation $\{F,G\}(y) = \nabla F(y)^T J^{-1}\nabla G(y)$, where $y=(p,q)$ and J is the matrix of (VI.2.3).

This Poisson bracket is bilinear, skew-symmetric ($\{F,G\} = -\{G,F\}$), it satisfies the *Jacobi identity* (Jacobi 1862, *Werke* 5, p. 46)

$$\big\{\{F,G\},H\big\} + \big\{\{G,H\},F\big\} + \big\{\{H,F\},G\big\} = 0 \tag{2.4}$$

(notice the cyclic permutations among F,G,H), and *Leibniz'* rule

$$\{F\cdot G,H\} = F\cdot\{G,H\} + G\cdot\{F,H\}. \tag{2.5}$$

These formulas are obtained in a straightforward manner from standard rules of calculus (see also Exercise 1).

With this notation, the Lie derivative (2.2) becomes

$$\frac{d}{dt}\,F\big(y(t)\big) \;=\; \{F,H\}(y(t)). \tag{2.6}$$

It follows that a function $I(p,q)$ is a first integral of (2.1) if and only if

$$\{I,H\} = 0.$$

If we take $F(y) = y_i$, the mapping that selects the ith component of y, we see that the Hamiltonian system (2.1) or (VI.2.5), $\dot{y} = J^{-1}\nabla H(y)$, can be written as

$$\dot{y}_i = \{y_i,H\}, \quad i = 1,\ldots,2d. \tag{2.7}$$

Poisson's Discovery. At the beginning of the 19th century, the hope of being able to integrate a given system of differential equations by analytic formulas faded more and more, and the energy of researchers went to the construction of, at least, first integrals. In this enthusiasm, Jacobi declared the subsequent result to be "Poisson's deepest discovery" (see citation) and his own identity, developed for its proof, a "gravissimum Theorema".



Siméon Denis Poisson[1]

Theorem 2.2 (Poisson 1809). *If I_1 and I_2 are first integrals, then their Poisson bracket $\{I_1, I_2\}$ is again a first integral.*

Proof. This follows at once from the Jacobi identity with $F = I_1$ and $G = I_2$. □

VII.2.2 General Poisson Structures

> ... the general concept of a Poisson manifold should be credited to Sophus Lie in his treatise on transformation groups ...
> (J.E. Marsden & T.S. Ratiu 1999)

We now come to the announced generalization of Definition 2.1 of the canonical Poisson bracket, invented by Lie (1888). Indeed, many proofs of properties of Hamiltonian systems rely uniquely on the bilinearity, the skew-symmetry and the Jacobi identity of the Poisson bracket, but not on the special structure of (2.3). So the idea is, more generally, to start with a smooth matrix-valued function $B(y) = \big(b_{ij}(y)\big)$ and to set

$$\{F, G\}(y) \;=\; \sum_{i,j=1}^{n} \frac{\partial F(y)}{\partial y_i}\, b_{ij}(y)\, \frac{\partial G(y)}{\partial y_j} \tag{2.8}$$

(or more compactly $\{F, G\}(y) = \nabla F(y)^T B(y) \nabla G(y)$).

Lemma 2.3. *The bracket defined in (2.8) is bilinear, skew-symmetric and satisfies Leibniz' rule (2.5) as well as the Jacobi identity (2.4) if and only if*

$$b_{ij}(y) = -b_{ji}(y) \qquad \text{for all } i, j \tag{2.9}$$

and for all i, j, k (notice the cyclic permutations among i, j, k)

$$\sum_{l=1}^{n} \Big(\frac{\partial b_{ij}(y)}{\partial y_l}\, b_{lk}(y) + \frac{\partial b_{jk}(y)}{\partial y_l}\, b_{li}(y) + \frac{\partial b_{ki}(y)}{\partial y_l}\, b_{lj}(y) \Big) \;=\; 0\,. \tag{2.10}$$

[1] Siméon Denis Poisson, born: 21 June 1781 in Pithiviers (France), died: 25 April 1840 in Sceaux (near Paris).

Proof. The main observation is that condition (2.10) is the Jacobi identity for the special choice of functions $F = y_i$, $G = y_j$, $H = y_k$ because of

$$\{y_i, y_j\} = b_{ij}(y). \tag{2.11}$$

If equation (2.4) is developed for the bracket (2.8), one obtains terms containing second order partial derivatives – these cancel due to the symmetry of the Jacobi identity – and terms containing first order partial derivatives; for the latter we may assume F, G, H to be linear combinations of y_i, y_j, y_k, so we are back to (2.10). The details of this proof are left as an exercise (see Exercise 1). □

Definition 2.4. If the matrix $B(y)$ satisfies the properties of Lemma 2.3, formula (2.8) is said to represent a (general) *Poisson bracket*. The corresponding differential system

$$\dot{y} = B(y)\nabla H(y), \tag{2.12}$$

is a *Poisson system*. We continue to call H a Hamiltonian.

The system (2.12) can again be written in the bracket formulation (2.7). The formula (2.6) for the Lie derivative remains also valid, as is seen immediately from the chain rule and the definition of the Poisson bracket. Choosing $F = H$, this shows in particular that the Hamiltonian H is a first integral for general Poisson systems.

Definition 2.5. A function $C(y)$ is called a *Casimir function* of the Poisson system (2.12), if

$$\nabla C(y)^T B(y) = 0 \qquad \text{for all } y.$$

A Casimir function is a first integral of every Poisson system with structure matrix $B(y)$, whatever the Hamiltonian $H(y)$ is.

Example 2.6. The *Lotka–Volterra* equations of Sect. I.1.1 can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & uv \\ -uv & 0 \end{pmatrix} \nabla H(u,v), \tag{2.13}$$

where $H(u,v) = u - \ln u + v - 2\ln v$ is the invariant (I.1.4). This is of the form (2.12) with a matrix that is skew-symmetric and satisfies the identity (2.10).

Higher dimensional Lotka–Volterra systems can also have a Poisson structure (see, e.g., Perelomov (1995) and Suris (1999)). For example, the system

$$\dot{y}_1 = y_1(y_2 + y_3), \quad \dot{y}_2 = y_2(y_1 - y_3 + 1), \quad \dot{y}_3 = y_3(y_1 + y_2 + 1)$$

can be written as

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & y_1y_2 & y_1y_3 \\ -y_1y_2 & 0 & -y_2y_3 \\ -y_1y_3 & y_2y_3 & 0 \end{pmatrix} \nabla H(y) \tag{2.14}$$

with $H(y) = -y_1 + y_2 + y_3 + \ln y_2 - \ln y_3$. Again one can check by direct computation that (2.10) is satisfied.

In contrast to the structure matrix J^{-1} of Hamiltonian systems in canonical form, the matrix $B(y)$ of (2.12) need not be invertible. All odd-dimensional skew-symmetric matrices are singular, and so is the matrix $B(y)$ of (2.14). In this case, the vector $v(y) = (-1/y_1, -1/y_2, 1/y_3)^T$ satisfies $v(y)^T B(y) = 0$. Since $v(y) = \nabla C(y)$ with $C(y) = -\ln y_1 - \ln y_2 + \ln y_3$, the function $C(y)$ is a Casimir function.

VII.2.3 Hamiltonian Systems on Symplectic Submanifolds

An important motivation for studying Poisson systems is given by Hamiltonian problems expressed in non-canonical coordinates.

Example 2.7 (Constrained Mechanical Systems). Consider the system (1.9) written as the differential equation

$$\dot{x} \;=\; J^{-1}\Big(\nabla H(x) + \sum_{i=1}^{m} \lambda_i(x)\, \nabla g_i(x)\Big) \tag{2.15}$$

on the manifold $\mathcal{M} = \{x\,;\, c(x) = 0\}$ with $c(x) = \big(g(q), G(q)H_p(p,q)\big)^T$ and $x = (p,q)^T$ (see (1.14)). As in the proof of Theorem 1.2, $\lambda_i(x)$ and $g_i(x)$ are the components of $\lambda(x)$ and $g(x)$, and $\lambda(x)$ is the function obtained from (1.12). We use $y \in \mathbb{R}^{2(d-m)}$ as local coordinates of the manifold $\mathcal{M}$ via the transformation

$$x = \chi(y).$$

In these coordinates, the differential equation (2.15) becomes, with $X(y) = \chi'(y)$,

$$X(y)\,\dot{y} \;=\; J^{-1}\Big(\nabla H\big(\chi(y)\big) + \sum_{i=1}^{m} \lambda_i\big(\chi(y)\big)\, \nabla g_i\big(\chi(y)\big)\Big).$$

We multiply this equation from the left with $X(y)^T J$ and note that the columns of $X(y)$, which are tangent vectors, are orthogonal to the gradients ∇g_i of the constraints. This yields

$$X(y)^T J X(y)\,\dot{y} \;=\; X(y)^T \nabla H\big(\chi(y)\big).$$

By assumption (1.13) the matrix $X(y)^T J X(y)$ is invertible. This is seen as follows: $X(y)^T J X(y) v = 0$ implies $J X(y) v = c'(x)^T w$ for some w ($x = \chi(y)$). By $c\big(\chi(y)\big) = 0$ and $c'(x)X(y) = 0$ we get $c'(x) J^{-1} c'(x)^T w = 0$. It then follows from the structure of $c'(x)$ and from (1.13) that $w = 0$ and hence also $v = 0$.

With $B(y) = \big(X(y)^T J X(y)\big)^{-1}$ and $K(y) = H\big(\chi(y)\big)$, the above equation for $\dot{y}$ thus becomes the Poisson system $\dot{y} = B(y)\nabla K(y)$. The matrix $B(y)$ is skew-symmetric and satisfies (2.10), see Theorem 2.8 below or Exercise 11.

More generally, consider a *symplectic submanifold* $\mathcal{M}$ of $\mathbb{R}^{2d}$, that is, a manifold for which the symplectic two-form[2]

$$\omega_x(\xi_1,\xi_2) = (J\xi_1,\xi_2) \qquad \text{for} \quad \xi_1,\xi_2 \in T_x\mathcal{M} \tag{2.16}$$

(with $(\cdot,\cdot)$ denoting the Euclidean inner product on $\mathbb{R}^{2d}$) is *non-degenerate* for every $x \in \mathcal{M}$: for ξ_1 in the tangent space $T_x\mathcal{M}$,

$$\omega_x(\xi_1,\xi_2) = 0 \;\text{ for all }\; \xi_2 \in T_x\mathcal{M} \quad \text{implies} \quad \xi_1 = 0.$$

In local coordinates $x = \chi(y)$, this condition is equivalent to the invertibility of the matrix $X(y)^T JX(y)$ with $X(y) = \chi'(y)$, since every tangent vector at $x = \chi(y)$ is of the form $\xi = X(y)\eta$ and $X(y)$ has linearly independent columns. A manifold defined by constraints, $\mathcal{M} = \{x \in \mathbb{R}^{2d} \,|\, c(x) = 0\}$, is symplectic if the matrix $c'(x)J^{-1}c'(x)^T$ is invertible for every $x \in \mathcal{M}$ (see the argument at the end of the previous example). This condition can be restated as saying that the matrix $\big(\{c_i,c_j\}(x)\big)$ of canonical Poisson brackets of the constraint functions is invertible.

We consider the reduction of the Hamiltonian system to the symplectic submanifold $\mathcal{M}$, which determines solution curves $t \mapsto x(t) \in \mathcal{M}$ by the equations

$$\big(J\dot{x} - \nabla H(x),\xi\big) = 0 \qquad \text{for all }\; \xi \in T_x\mathcal{M}. \tag{2.17}$$

With the interpretation $(\nabla H(x),\xi) = H'(x)\xi = \frac{d}{dt}\big|_{t=0} H(\gamma(t))$ as a directional derivative along a path $\gamma(t) \in \mathcal{M}$ with $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi$, it is sufficient that the Hamiltonian H is defined and differentiable on the manifold $\mathcal{M}$. Equation (2.17) can also be expressed as

$$\omega_x(\dot{x},\xi) = H'(x)\xi \qquad \text{for all }\; \xi \in T_x\mathcal{M}, \tag{2.18}$$

a formulation that is susceptible to further generalization; cf. Marsden & Ratiu (1999), Chap. 5.4, and Exercise 2. Choosing $\xi = \dot{x}$ we obtain $0 = H'(x)\dot{x} = \frac{d}{dt}H(x(t))$, and hence the Hamiltonian is conserved along solutions.

Note that for $\mathcal{M}$ of Example 2.7, the formulation (2.17) is equivalent to the equations of motion (2.15) of the constrained mechanical system. It corresponds to *d'Alembert's principle of virtual variations* in constrained mechanics; see Arnold (1989), p. 92. In quantum mechanics the Hamiltonian reduction (2.17) to a manifold (in that case, a submanifold of the Hilbert space $L^2(\mathbb{R}^N,\mathbb{R}^2)$ instead of $\mathbb{R}^{2d}$) is known as the *Dirac–Frenkel time-dependent variational principle* and is the basic tool for deriving reduced models of the many-body Schrödinger equation; see Sect. VII.6 for an example. From a numerical analysis viewpoint, (2.17) can also be viewed as a Galerkin method on the solution-dependent tangent space $T_x\mathcal{M}$.

In terms of the *symplectic projection* $P(x) : \mathbb{R}^{2d} \to T_x\mathcal{M}$ for $x \in \mathcal{M}$, defined by determining $P(x)v \in T_x\mathcal{M}$ for $v \in \mathbb{R}^{2d}$ from the condition

$$(JP(x)v,\xi) = (Jv,\xi) \qquad \text{for all} \quad \xi \in T_x\mathcal{M}, \tag{2.19}$$

[2] Notice that this two-form is the negative of that introduced in Sect. VI.2. This slight inconsistency makes the subsequent formulas nicer.

formula (2.17) can be reformulated as the differential equation on $\mathcal{M}$,

$$\dot{x} = P(x)J^{-1}\nabla H(x). \tag{2.20}$$

In coordinates $x = \chi(y)$, and again with $X(y) = \chi'(y)$, formula (2.17) becomes

$$X(y)^T \Big(JX(y)\dot{y} - \nabla H(\chi(y)) \Big) = 0,$$

and with

$$B(y) = \big(X(y)^T JX(y) \big)^{-1} \qquad \text{and} \qquad K(y) = H(\chi(y)), \tag{2.21}$$

we obtain the differential equation

$$\dot{y} = B(y)\nabla K(y). \tag{2.22}$$

Theorem 2.8. *For a Hamiltonian system (2.17) on a symplectic submanifold $\mathcal{M}$, the equivalent differential equation in local coordinates, (2.22) with (2.21), is a Poisson system.*

Proof. In coordinates, the symplectic projection is given by

$$P(x) = X(y)B(y)X(y)^T J \qquad \text{for} \quad x = \chi(y) \in \mathcal{M},$$

since for every tangent vector $\xi = X(y)\eta$ we have by (2.21),

$$(JXBX^T Jv, X\eta) = (X^T JXBX^T Jv, \eta) = (X^T Jv, \eta) = (Jv, X\eta).$$

From the decomposition $\mathbb{R}^{2d} = P(x)\mathbb{R}^{2d} \oplus (I - P(x))\mathbb{R}^{2d}$ we obtain, by the implicit function theorem, a corresponding splitting in a neighbourhood of the manifold $\mathcal{M}$ in $\mathbb{R}^{2d}$,

$$v = x + w \quad \text{with} \ x \in \mathcal{M}, \ P(x)w = 0.$$

This permits us to extend smooth functions $F(y)$ to a neighbourhood of $\mathcal{M}$ by setting

$$\widehat{F}(v) = F(y) \quad \text{for} \ v = x + w \ \text{with} \ x = \chi(y), \ P(x)w = 0.$$

We then have for the derivative $\widehat{F}'(x) = \widehat{F}'(x)P(x)$ for $x \in \mathcal{M}$ and hence for its transpose, the gradient, $\nabla\widehat{F}(x) = P(x)^T \nabla\widehat{F}(x)$. Moreover, by the chain rule we have $\nabla F(y) = X(y)^T \nabla\widehat{F}(x)$ for $x = \chi(y)$. For the canonical bracket this gives, at $x = \chi(y)$,

$$\begin{aligned} \{\widehat{F}, \widehat{G}\}_{\text{can}}(x) &= \nabla\widehat{F}(x)^T P(x)J^{-1}P(x)^T \nabla\widehat{G}(x) \\ &= \nabla F(y)^T B(y)\nabla G(y) \ = \ \{F, G\}(y), \end{aligned}$$

and hence the required properties of the bracket defined by $B(y)$ follow from the corresponding properties of the canonical bracket. □

VII.3 The Darboux–Lie Theorem

Theorem 2.8 also shows that a Hamiltonian system without constraints becomes a Poisson system in non-canonical coordinates. Interestingly, a converse also holds: every Poisson system can locally be written in canonical Hamiltonian form after a suitable change of coordinates. This result is a special case of the *Darboux–Lie Theorem*. Its proof was the result of several important papers: Jacobi's theory of simultaneous linear partial differential equations (Jacobi 1862), the works by Clebsch (1866) and Darboux (1882) on Pfaffian systems, and, finally, the paper of Lie (1888). We shall now retrace this development. Our first tool is a result on the commutativity of Poisson flows.

VII.3.1 Commutativity of Poisson Flows and Lie Brackets

The elegant formula (2.6) for the Lie derivative is valid for general Poisson systems with the vector field $f(y) = B(y)\nabla H(y)$ of (2.12). Acting on a function $F : \mathbb{R}^n \to \mathbb{R}$, the Lie operator (III.5.2) becomes

$$DF = \nabla F^T f = \nabla F^T B(y) \nabla H = \{F, H\} \tag{3.1}$$

and is again the Poisson bracket. This observation is the key for the following lemma, which shows an interesting connection between the Lie bracket and the Poisson bracket.

Lemma 3.1. *Let two smooth Hamiltonians $H^{[1]}(y)$ and $H^{[2]}(y)$ be given.*

$$\begin{array}{cccc} \textit{If} & D_1 & \textit{is the Lie operator of} & B(y)\nabla H^{[1]} \\ \textit{and} & D_2 & \textit{is the Lie operator of} & B(y)\nabla H^{[2]}, \\ \textit{then} & [D_1, D_2] & \textit{is the Lie operator of} & B(y)\nabla \{H^{[2]}, H^{[1]}\} \end{array} \tag{3.2}$$

(notice, once again, that the indices 1 *and* 2 *have been reversed).*

Proof. After some clever permutations, the Jacobi identity (2.4) can be written as

$$\big\{\{F, H^{[2]}\}, H^{[1]}\big\} - \big\{\{F, H^{[1]}\}, H^{[2]}\big\} = \big\{F, \{H^{[2]}, H^{[1]}\}\big\}. \tag{3.3}$$

By (3.1) this is nothing other than $D_1 D_2 F - D_2 D_1 F = [D_1, D_2]F$. □

Lemma 3.2. *Consider two smooth Hamiltonians $H^{[1]}(y)$ and $H^{[2]}(y)$ on an open connected set U, with D_1 and D_2 the corresponding Lie operators and $\varphi_s^{[1]}(y)$ and $\varphi_t^{[2]}(y)$ the corresponding flows. Then, if the matrix $B(y)$ is invertible, the following are equivalent in U:*

$$\begin{array}{rl} \textit{(i)} & \{H^{[1]}, H^{[2]}\} = \textit{Const}; \\ \textit{(ii)} & [D_1, D_2] = 0; \\ \textit{(iii)} & \varphi_t^{[2]} \circ \varphi_s^{[1]} = \varphi_s^{[1]} \circ \varphi_t^{[2]}. \end{array}$$

The conclusions "(i) ⇒ (ii) ⇔ (iii)" also hold for a non-invertible $B(y)$.

Proof. This is obtained by combining Lemma III.5.4 and Lemma 3.1. We need the invertibility of $B(y)$ to conclude that $\{H^{[1]}, H^{[2]}\} = Const$ follows from $B(y)\nabla\{H^{[1]}, H^{[2]}\} = 0$. □

VII.3.2 Simultaneous Linear Partial Differential Equations

If two functions $F(y)$ and $G(y)$ are given, formula (2.8) determines a function $h(y) = \{F, G\}(y)$ by differentiation. We now ask the *inverse* question: Given functions $G(y)$ and $h(y)$, can we find a function $F(y)$ such that $\{F, G\}(y) = h(y)$? This problem represents a first order linear partial differential equation for F. So we are led to the following problem, which we first discuss in two dimensions.

One Equation. Given functions $a(y_1, y_2)$, $b(y_1, y_2)$, $h(y_1, y_2)$, find all solutions $F(y_1, y_2)$ satisfying

$$a(y_1, y_2)\,\frac{\partial F}{\partial y_1} + b(y_1, y_2)\,\frac{\partial F}{\partial y_2} = h(y_1, y_2). \tag{3.4}$$

This equation is, for any point (y_1, y_2), a linear relation between the partial derivatives of F, but does not determine them individually. There is *one* direction, however, where the derivative is uniquely determined, namely that of the vector $\boldsymbol{n} = \big(a(y_1, y_2), b(y_1, y_2)\big)$, since the left-hand side of equation (3.4) is the directional derivative $\frac{\partial F}{\partial \boldsymbol{n}}$. The lines, which everywhere respect this direction, are called *characteristic lines* (see left picture of Fig. 3.1). If we parametrize them with a parameter t, we can compute $y_1(t)$, $y_2(t)$ as well as $F(t) = F\big(y_1(t), y_2(t)\big)$ as solutions of the following ordinary differential equations

$$\dot y_1 = a(y_1, y_2), \qquad \dot y_2 = b(y_1, y_2), \qquad \dot F = h(y_1, y_2). \tag{3.5}$$

The *initial values* $\big(y_1(0), y_2(0)\big)$ can be chosen on an arbitrary curve γ (which must be transversal to the characteristic lines) and the values $F|_\gamma$ can be arbitrarily prescribed. The solution $F(y_1, y_2)$ of (3.4) is then created by the curves (3.5) wherever the characteristic lines go (right picture of Fig. 3.1).

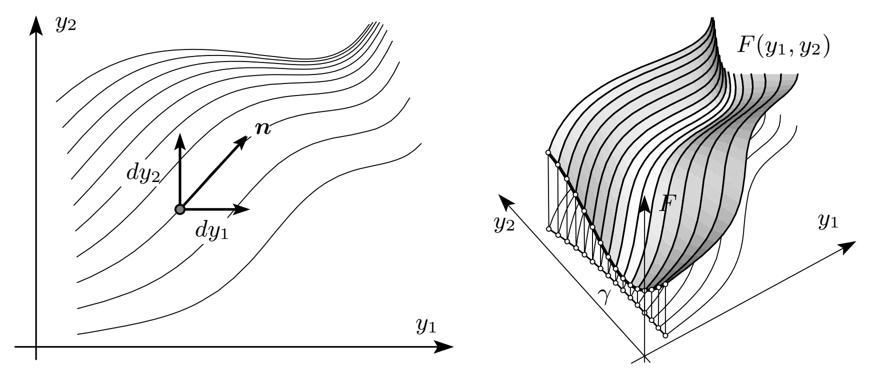


Fig. 3.1. Characteristic lines and solution of a first order linear partial differential equation

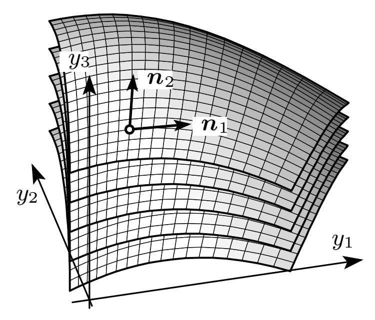


Fig. 3.2. Characteristic surfaces of two first order linear partial differential equations

For one equation in n dimensions, the initial values $\big(y_1(0),\dots,y_n(0)\big)$ can be freely chosen on a manifold of dimension $n-1$ (e.g., the subspace orthogonal to the characteristic line passing through a given point), and F can be arbitrarily prescribed on this manifold. This guarantees the existence of $n-1$ independent solutions in the neighbourhood of a given point. Here, independent means that the gradients of these functions are linearly independent.

Two Simultaneous Equations. Two simultaneous equations of dimension two are trivial. We therefore suppose $y=(y_1,y_2,y_3)$ and two equations of the form

$$\begin{aligned} a_1^{[1]}(y)\,\frac{\partial F}{\partial y_1}+a_2^{[1]}(y)\,\frac{\partial F}{\partial y_2}+a_3^{[1]}(y)\,\frac{\partial F}{\partial y_3}&=h_1(y),\\ a_1^{[2]}(y)\,\frac{\partial F}{\partial y_1}+a_2^{[2]}(y)\,\frac{\partial F}{\partial y_2}+a_3^{[2]}(y)\,\frac{\partial F}{\partial y_3}&=h_2(y)\end{aligned} \tag{3.6}$$

for an unknown function $F(y_1,y_2,y_3)$. This system can also be written as $D_1F=h_1, D_2F=h_2$, where D_i denotes the Lie operator corresponding to the vector field $a^{[i]}(y)$. Here, we have *two* directional derivatives prescribed, namely $\frac{\partial F}{\partial \boldsymbol{n}_1}$ and $\frac{\partial F}{\partial \boldsymbol{n}_2}$ where $\boldsymbol{n}_i=a^{[i]}(y)$ (see Fig. 3.2). Therefore, we will have to follow both directions and, instead of (3.5), we will have *two* sets of ordinary differential equations

$$\begin{aligned} &\dot y_1=a_1^{[1]}(y),\qquad \dot y_2=a_2^{[1]}(y),\qquad \dot y_3=a_3^{[1]}(y),\qquad \dot F=h_1(y)\\ &\dot y_1=a_1^{[2]}(y),\qquad \dot y_2=a_2^{[2]}(y),\qquad \dot y_3=a_3^{[2]}(y),\qquad \dot F=h_2(y).\end{aligned} \tag{3.7}$$

If we prescribe F on a curve that is orthogonal to $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$, and if we follow the solutions of (3.7), we obtain the function F on two 2-dimensional surfaces S_1 and S_2 containing the prescribed curve. Continuing from S_1 along the second flow and from S_2 along the first flow, we may be led to the same point, but nothing guarantees that the obtained values for F are identical. To get a well-defined F, additional assumptions on the differential operators and on the inhomogeneities have to be made.

The following theorem, which is due to Jacobi (1862), has been extended by Clebsch (1866), who created the theory of *complete systems* (“vollständige

Systeme”). These papers contained long analytic calculations with myriades of formulas. The wonderful geometric insight is mainly due to Sophus Lie.

Theorem 3.3. *Let $D_1,\ldots,D_m$ be m $(m<n)$ linear differential operators in $\mathbb{R}^n$ corresponding to vector fields $a^{[1]}(y),\ldots,a^{[m]}(y)$ and suppose that these vectors are linearly independent for $y=y_0$. If*

$$[D_i,D_j]=0 \qquad \text{for all } i,j, \tag{3.8}$$

then the homogeneous system

$$D_iF=0 \qquad \text{for } \; i=1,\ldots,m$$

possesses (in a neighbourhood of y_0) $n-m$ solutions for which the gradients $\nabla F(y_0)$ are linearly independent.

Furthermore, the inhomogeneous system of partial differential equations

$$D_iF=h_i \qquad \text{for } \; i=1,\ldots,m$$

possesses a particular solution in a neighbourhood of y_0, if and only if in addition to (3.8) the functions $h_1(y),\ldots,h_m(y)$ satisfy the integrability conditions

$$D_ih_j=D_jh_i \qquad \text{for all } i,j. \tag{3.9}$$

Proof. (a) Let V denote the space of vectors in $\mathbb{R}^n$ that are orthogonal to $a^{[1]}(y_0)$, $\ldots,a^{[m]}(y_0)$, and consider the $(n-m)$-dimensional manifold $\mathcal{M}=y_0+V$. We then extend an arbitrary smooth function $F:\mathcal{M}\to\mathbb{R}$ to a neighbourhood of y_0 by

$$F\Big(\varphi^{[m]}_{t_m}\circ\ldots\circ\varphi^{[1]}_{t_1}(y_0+v)\Big)=F\big(y_0+v\big). \tag{3.10}$$

Notice that $(t_1,\ldots,t_m,v)\mapsto y=\varphi^{[m]}_{t_m}\circ\ldots\circ\varphi^{[1]}_{t_1}(y_0+v)$ defines a local diffeomorphism between neighbourhoods of 0 and y_0. Since the application of the operator D_m to (3.10) corresponds to a differentiation with respect to t_m and the expression $F\big(\varphi^{[m]}_{t_m}\circ\ldots\circ\varphi^{[1]}_{t_1}(y_0+v)\big)$ is independent of t_m by (3.10), we get $D_mF(y)=0$. To prove $D_iF(y)=0$ for $i<m$, we first have to change the order of the flows $\varphi^{[j]}_{t_j}$ in (3.10), which is permitted by Lemma III.5.4 and assumption (3.8), so that $\varphi^{[i]}_{t_i}$ is in the left-most position.

(b) The necessity of (3.9) follows immediately from $D_ih_j=D_iD_jF=D_jD_iF=D_jh_i$. For given h_i satisfying (3.9) we define $F(y)$ in a neighbourhood of y_0 (i.e., for small $t_1,\ldots,t_m$ and small v) by

$$\begin{aligned}F\Big(\varphi^{[m]}_{t_m}\circ\ldots\circ\varphi^{[1]}_{t_1}(y_0+v)\Big) \;&=\; \int_0^{t_1} h_1\Big(\varphi^{[1]}_{t}(y_0+v)\Big)\,dt\\ &+\ldots+\int_0^{t_m} h_m\Big(\varphi^{[m]}_{t}\circ\varphi^{[m-1]}_{t_{m-1}}\circ\ldots\circ\varphi^{[1]}_{t_1}(y_0+v)\Big)\,dt,\end{aligned}$$

and we prove that it is a solution of the system $D_iF = h_i$ for $i = 1, \ldots, m$. Since only the last integral depends on t_m, we immediately get by differentiation with respect to t_m that $D_mF = h_m$. For the computation of D_iF we differentiate with respect to t_i. The first $i-1$ integrals are independent of t_i. The derivative of the ith integral gives $h_i\big(\varphi_{t_i}^{[i]} \circ \ldots \circ \varphi_{t_1}^{[1]}(y_0+v)\big)$, and the derivative of the remaining integrals gives

$$\int_0^{t_j} D_i h_j\Big(\varphi_t^{[j]} \circ \ldots \circ \varphi_{t_1}^{[1]}(y_0+v)\Big)\,dt = \int_0^{t_j} D_j h_i\Big(\varphi_t^{[j]} \circ \ldots \circ \varphi_{t_1}^{[1]}(y_0+v)\Big)\,dt$$
$$= h_i\Big(\varphi_{t_j}^{[j]} \circ \ldots \circ \varphi_{t_1}^{[1]}(y_0+v)\Big) - h_i\Big(\varphi_{t_{j-1}}^{[j-1]} \circ \ldots \circ \varphi_{t_1}^{[1]}(y_0+v)\Big)$$

for $j = i+1, \ldots, m$. Summing up, this proves $D_iF = h_i$. □

VII.3.3 Coordinate Changes and the Darboux–Lie Theorem

The emphasis here is to simplify a given Poisson structure as much as possible by a coordinate transformation. We change from coordinates $y_1, \ldots, y_n$ to $\widetilde{y}_1(y), \ldots, \widetilde{y}_n(y)$ with continuously differentiable functions and an invertible Jacobian $A(y) = \partial\widetilde{y}/\partial y$,

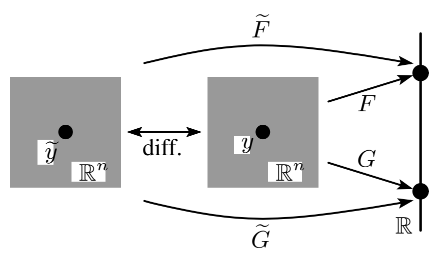




Fig. 3.3. New coordinates in a Poisson system



Jean Gaston Darboux[3]

and we denote $\widetilde{F}(\widetilde{y}) := F(y)$ and $\widetilde{G}(\widetilde{y}) := G(y)$ (see Fig. 3.3). The Poisson structure as well as the Poisson flow on one space will become another Poisson structure and flow on the other space by simply applying the chain rule:

$$\sum_{i,j} \frac{\partial F(y)}{\partial y_i}\, b_{ij}(y)\, \frac{\partial G(y)}{\partial y_j} = \sum_{i,j,k,l} \frac{\partial \widetilde{F}(\widetilde{y})}{\partial \widetilde{y}_k}\, \frac{\partial \widetilde{y}_k}{\partial y_i}\, b_{ij}\big(y(\widetilde{y})\big)\, \frac{\partial \widetilde{y}_l}{\partial y_j}\, \frac{\partial \widetilde{G}(\widetilde{y})}{\partial \widetilde{y}_l}. \tag{3.11}$$

This is another Poisson structure with

$$\widetilde{b}_{kl} = \{\widetilde{y}_k, \widetilde{y}_l\} \qquad \text{or} \qquad \widetilde{B}(\widetilde{y}) = A(y)B(y)A(y)^T. \tag{3.12}$$

[3] Jean Gaston Darboux, born: 14 August 1842 in Nîmes (France), died: 23 February 1917 in Paris.

The same structure matrix is obtained if the Poisson system (2.12) is written in these new coordinates (Exercise 5).

Since A is invertible, the structure matrices B and $\widetilde{B}$ have the same rank. We now want to obtain the simplest possible form for $\widetilde{B}$.

Theorem 3.4 (Darboux 1882, Lie 1888). *Suppose that the matrix $B(y)$ defines a Poisson bracket and is of constant rank $n-q=2m$ in a neighbourhood of $y_0 \in \mathbb{R}^n$. Then, there exist functions $P_1(y),\ldots,P_m(y)$, $Q_1(y),\ldots,Q_m(y)$, and $C_1(y),\ldots,C_q(y)$ satisfying*

$$\begin{array}{lll} \{P_i,P_j\}=0 & \{P_i,Q_j\}=-\delta_{ij} & \{P_i,C_l\}=0 \\ \{Q_i,P_j\}=\delta_{ij} & \{Q_i,Q_j\}=0 & \{Q_i,C_l\}=0 \\ \{C_k,P_j\}=0 & \{C_k,Q_j\}=0 & \{C_k,C_l\}=0 \end{array} \tag{3.13}$$

on a neighbourhood of y_0. The gradients of P_i,Q_i,C_k are linearly independent, so that $y \mapsto \big(P_i(y),Q_i(y),C_k(y)\big)$ constitutes a local change of coordinates to canonical form.

The functions $C_1(y),\ldots,C_q(y)$ are called *distinguished functions* (ausgezeichnete Funktionen) by Lie.

Proof. We follow Lie's original proof. Similar ideas, and the same notation, are also present in Darboux's paper. The proof proceeds in several steps, satisfying the conditions of (3.13), from one line to the next, by solving systems of linear partial differential equations.

(a) If all $b_{ij}(y_0)=0$, the constant rank assumption implies $b_{ij}(y)=0$ in a neighbourhood of y_0. We thus have $m=0$ and all coordinates $C_i(y)=y_i$ are Casimirs.

(b) If there exist i,j with $b_{ij}(y_0)\neq 0$, we set $Q_1(y)=y_i$ and we determine $P_1(y)$ as the solution of the linear partial differential equation

$$\{Q_1,P_1\}=1. \tag{3.14}$$

Because of $b_{ij}(y_0)\neq 0$ the assumption of Theorem 3.3 is satisfied and this yields the existence of P_1. We next consider the homogeneous system

$$\{Q_1,F\}=0 \qquad \text{and} \qquad \{P_1,F\}=0 \tag{3.15}$$

of partial differential equations. By Lemma 3.2 and (3.14) the Lie operators corresponding to Q_1 and P_1 commute, so that by Theorem 3.3 the system (3.15) has $n-2$ independent solutions $F_3,\ldots,F_n$. Their gradients together with those of Q_1 and P_1 form a basis of $\mathbb{R}^n$. We therefore can change coordinates from $y_1,\ldots,y_n$ to $Q_1,P_1,F_3,\ldots,F_n$ (mapping y_0 to $\widetilde{y}_0$). In these coordinates the first two rows and the first two columns of the structure matrix $\widetilde{B}(\widetilde{y})$ have the required form.

(c) If $\widetilde{b}_{ij}(\widetilde{y}_0)=0$ for all $i,j\geq 3$, we have $m=1$ (similar to step (a)) and the coordinates $F_3,\ldots,F_n$ are Casimirs.

(d) If there exist $i \geq 3$ and $j \geq 3$ with $\widetilde{b}_{ij}(\widetilde{y}_0) \neq 0$, we set $Q_2 = F_i$ and we determine P_2 from the inhomogeneous system

$$\{Q_1, P_2\} = 0, \quad \{P_1, P_2\} = 0, \quad \{Q_2, P_2\} = 1.$$

The inhomogeneities satisfy (3.9), and the Lie operators corresponding to Q_1, P_1, Q_2 commute (by Lemma 3.2). Theorem 3.3 proves the existence of such a P_2. We then consider the homogeneous system

$$\{Q_1, F\} = 0, \quad \{P_1, F\} = 0, \quad \{Q_2, F\} = 0, \quad \{P_2, F\} = 0$$

and apply once more Theorem 3.3. We get $n-4$ independent solutions, which we denote again $F_5, \ldots, F_n$. As in part (b) of the proof we get new coordinates $Q_1, P_1, Q_2, P_2, F_5, \ldots, F_n$, for which the first *four* rows and columns of the structure matrix are canonical.

(e) The proof now continues by repeating steps (c) and (d) until the structure matrix has the desired form. □

Corollary 3.5 (Casimir Functions). *In the situation of Theorem 3.4 the functions* $C_1(y), \ldots, C_q(y)$ *satisfy*

$$\{C_i, H\} = 0 \qquad \textit{for all smooth } H. \tag{3.16}$$

Proof. Theorem 3.4 states that $\nabla C_i(y)^T B(y) \nabla H(y) = 0$, when $H(y)$ is one of the functions $P_j(y), Q_j(y)$ or $C_j(y)$. However, the gradients of these functions form a basis of $\mathbb{R}^n$. Consequently, $\nabla C_i(y)^T B(y) = 0$ and (3.16) is satisfied for all differentiable functions $H(y)$. □

This property implies that all Casimir functions are first integrals of (2.12) whatever $H(y)$ is. Consequently, (2.12) is (close to y_0) a differential equation on the manifold

$$\mathcal{M} = \{y \in U \mid C_i(y) = Const_i,\ i = 1, \ldots, m\}. \tag{3.17}$$

Corollary 3.6 (Transformation to Canonical Form). *Denote the transformation of Theorem 3.4 by* $z = \vartheta(y) = \big(P_i(y), Q_i(y), C_k(y)\big)$*. With this change of coordinates, the Poisson system* $\dot{y} = B(y)\nabla H(y)$ *becomes*

$$\dot{z} = B_0 \nabla K(z) \qquad \textit{with} \qquad B_0 = \begin{pmatrix} J^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.18}$$

where $K(z) = H(y)$*. Writing* $z = (p, q, c)$*, this system becomes*

$$\dot{p} = -K_q(p, q, c), \qquad \dot{q} = K_p(p, q, c), \qquad \dot{c} = 0.$$

Proof. The transformed differential equation is

$$\dot{z} = \vartheta'(y) B(y) \vartheta'(y)^T \nabla K(z) \qquad \text{with} \qquad y = \vartheta^{-1}(z),$$

and Theorem 3.4 states that $\vartheta'(y) B(y) \vartheta'(y)^T = B_0$. □

VII.4 Poisson Integrators

Before discussing geometric numerical integrators, we show that many important properties of Hamiltonian systems in canonical form remain valid for systems

$$\dot{y} = B(y)\nabla H(y), \tag{4.1}$$

where $B(y)$ represents a Poisson bracket.

VII.4.1 Poisson Maps and Symplectic Maps

We have already seen that the Hamiltonian $H(y)$ is a first integral of (4.1). We shall show here that the flow of (4.1) satisfies a property closely related to symplecticity.

Definition 4.1. A transformation $\varphi : U \to \mathbb{R}^n$ (where U is an open set in $\mathbb{R}^n$) is called a *Poisson map* with respect to the bracket (2.8), if its Jacobian matrix satisfies

$$\varphi'(y)B(y)\varphi'(y)^T \;=\; B\big(\varphi(y)\big). \tag{4.2}$$

An equivalent condition is that for all smooth real-valued functions F, G defined on $\varphi(U)$,

$$\{F\circ\varphi, G\circ\varphi\}(y) = \{F, G\}\big(\varphi(y)\big), \tag{4.3}$$

as is seen by the chain rule and choosing F, G as the coordinate functions. It is clear from this condition that the composition of Poisson maps is again a Poisson map. A comparison with (3.12) shows that Poisson maps leave the structure matrix invariant.

For the canonical symplectic structure, where $B(y) = J^{-1}$, condition (4.2) is equivalent to the symplecticity of the transformation $\varphi(y)$. This can be seen by taking the inverse of both sides of (4.2), and by multiplying the resulting equation with $\varphi'(y)$ from the right and with $\varphi'(y)^T$ from the left. Also in the situation of a Hamiltonian system (2.17) on a symplectic submanifold $\mathcal{M}$, where $B(y)$ is the structure matrix of the differential equation in coordinates y as in Theorem 2.8, condition (4.2) is equivalent to symplecticity in the sense of preserving the symplectic two-form (2.16) on the tangent space, as in (1.16):

Definition 4.2. A map $\psi : \mathcal{M} \to \mathcal{M}$ on a symplectic manifold $\mathcal{M}$ is called *symplectic* if for every $x \in \mathcal{M}$,

$$\omega_{\psi(x)}\big(\psi'(x)\xi_1, \psi'(x)\xi_2\big) = \omega_x(\xi_1, \xi_2) \qquad \text{for all} \quad \xi_1, \xi_2 \in T_x\mathcal{M}. \tag{4.4}$$

A near-identity map $\psi : \mathcal{M} \to \mathcal{M}$ is symplectic if and only if the conjugate map φ in local coordinates $x = \chi(y)$, with $\varphi(y)$ given by $\psi(x) = \chi(\varphi(y))$ for $x = \chi(y)$, is a Poisson map for the structure matrix of (2.21), $B(y) = (X(y)^T JX(y))^{-1}$ with $X(y) = \chi'(y)$. This holds because $\psi'(x)\xi = X(\varphi(y))\varphi'(y)\eta$ for $x = \chi(y)$ and $\xi = X(y)\eta$, and because (4.2) is equivalent to $\varphi'(y)^T X(\varphi(y))^T JX(\varphi(y))\varphi'(y) = X(y)^T JX(y)$.

Theorem 4.3. *If $B(y)$ is the structure matrix of a Poisson bracket, then the flow $\varphi_t(y)$ of the differential equation (4.1) is a Poisson map.*

Proof. (a) For $B(y) = J^{-1}$ this is exactly the statement of Theorem VI.2.4 on the symplecticity of the flow of Hamiltonian systems. This result can be extended in a straightforward way to the matrix B_0 of (3.18).

(b) For the general case consider the change of coordinates $z = \vartheta(y)$ which transforms (4.1) to canonical form (Theorem 3.4), i.e., $\vartheta'(y)B(y)\vartheta'(y)^T = B_0$ and $\dot{z} = B_0 \nabla K(z)$ with $K(z) = H(y)$ (Corollary 3.6). Denoting the flows of (4.1) and $\dot{z} = B_0 \nabla K(z)$ by $\varphi_t(y)$ and $\psi_t(z)$, respectively, we have $\psi_t\big(\vartheta(y)\big) = \vartheta\big(\varphi_t(y)\big)$ and by the chain rule $\psi_t'\big(\vartheta(y)\big)\vartheta'(y) = \vartheta'\big(\varphi_t(y)\big)\varphi_t'(y)$. Inserting this relation into $\psi_t'(z)B_0\psi_t'(z)^T = B_0$, which follows from (a), proves the statement.

A direct proof, avoiding the use of Theorem 3.4, is indicated in Exercise 6. □

From Theorems 2.8 and 4.3 and the remark after Definition 4.2 we note the following.

Corollary 4.4. *The flow of a Hamiltonian system (2.17) on a symplectic submanifold is symplectic.*

The inverse of Theorem 4.3 is also true. It extends Theorem VI.2.6 from canonically symplectic transformations to Poisson maps.

Theorem 4.5. *Let $f(y)$ and $B(y)$ be continuously differentiable on an open set $U \subset \mathbb{R}^m$, and assume that $B(y)$ represents a Poisson bracket (Definition 2.4). Then, $\dot{y} = f(y)$ is locally of the form (4.1), if and only if*

- *its flow $\varphi_t(y)$ respects the Casimirs of $B(y)$, i.e., $C_i\big(\varphi_t(y)\big) = Const$, and*
- *its flow is a Poisson map for all $y \in U$ and for all sufficiently small t.*

Proof. The necessity follows from Corollary 3.5 and from Theorem 4.3. For the proof of sufficiency we apply the change of coordinates $(u, c) = \vartheta(y)$ of Theorem 3.4, which transforms $B(y)$ into canonical form (3.18). We write the differential equation $\dot{y} = f(y)$ in the new variables as

$$\dot{u} = g(u, c), \qquad \dot{c} = h(u, c). \tag{4.5}$$

Our first assumption expresses the fact that the Casimirs, which are the components of c, are first integrals of this system. Consequently, we have $h(u, c) \equiv 0$. The second assumption implies that the flow of (4.5) is a Poisson map for B_0 of (3.18). Writing down explicitly the blocks of condition (4.2), we see that this is equivalent to the symplecticity of the mapping $u_0 \mapsto u(t, u_0, c_0)$, with c_0 as a parameter. From Theorem VI.2.6 we thus obtain the existence of a function $K(u, c)$ such that $g(u, c) = J^{-1}\nabla_u K(u, c)$. Notice that for flows depending smoothly on a parameter, the Hamiltonian also depends smoothly on it. Consequently, the vector field (4.5) is of the form $B_0 \nabla K(u, c)$. Transforming back to the original variables we obtain $f(y) = B(y)\nabla H(y)$ with $H(y) = K\big(\vartheta(y)\big)$ (see Corollary 3.6). □

VII.4.2 Poisson Integrators

The preceding theorem shows that "being a Poisson map and respecting the Casimirs" is characteristic for the flow of a Poisson system. This motivates the following definition.

Definition 4.6. A numerical method $y_1 = \Phi_h(y_0)$ is a *Poisson integrator* for the structure matrix $B(y)$, if the transformation $y_0 \mapsto y_1$ respects the Casimirs and if it is a Poisson map whenever the method is applied to (4.1).

Observe that for a Poisson integrator one has to specify the class of structure matrices $B(y)$. A method will never be a Poisson integrator for all possible $B(y)$.

Example 4.7. The symplectic Euler method reads

$$u_{n+1} = u_n + hu_{n+1}v_n H_v(u_{n+1}, v_n), \quad v_{n+1} = v_n - hu_{n+1}v_n H_u(u_{n+1}, v_n)$$

for the Lotka–Volterra problem (2.13). It produces an excellent long-time behaviour (Fig. 4.1, left picture). We shall show that this is a Poisson integrator for all separable Hamiltonians $H(u, v) = K(u)+L(v)$. For this we compute the Jacobian of the map $(u_n, v_n) \mapsto (u_{n+1}, v_{n+1})$,

$$\begin{pmatrix} 1 - hv_n H_v & 0 \\ hv_n(H_u+u_{n+1}H_{uu}) & 1 \end{pmatrix} \begin{pmatrix} \dfrac{\partial(u_{n+1}, v_{n+1})}{\partial(u_n, v_n)} \end{pmatrix} = \begin{pmatrix} 1 & hu_{n+1}(H_v+v_n H_{vv}) \\ 0 & 1 - hu_{n+1}H_u \end{pmatrix}$$

(the argument of the partial derivatives of H is (u_{n+1}, v_n) everywhere), and we check in a straightforward fashion the validity of (4.2). A different proof, using differential forms, is given in Sanz-Serna (1994) for a special choice of $H(u, v)$. Similarly, the adjoint of the symplectic Euler method is a Poisson integrator, and so is their composition – the Störmer–Verlet scheme. Composition methods based on this scheme yield high order Poisson integrators, because the composition of Poisson maps is again a Poisson map.

The implicit midpoint rule, though symplectic when applied to canonical Hamiltonian systems, turns out not to be a Poisson map for the structure matrix $B(u, v)$ of (2.13). Figure 4.1 (right picture) shows that the numerical solution does not remain near a closed curve.

It is a difficult task to construct Poisson integrators for general Poisson systems; cf. the overview by Karasözen (2004). First of all, for non-constant $B(y)$ condition (4.2) is no longer a quadratic first integral of the problem augmented by its variational equation (see Sect. VI.4.1). Secondly, the Casimir functions can be arbitrary and we know that only linear and quadratic first integrals can be conserved automatically (Chap. IV). Therefore, Poisson integrators will have to exploit special structures of the particular problem.

Splitting Methods. Consider a (general) Poisson system $\dot{y} = B(y)\nabla H(y)$ and suppose that the Hamiltonian permits a decomposition as $H(y) = H^{[1]}(y) + \ldots +$

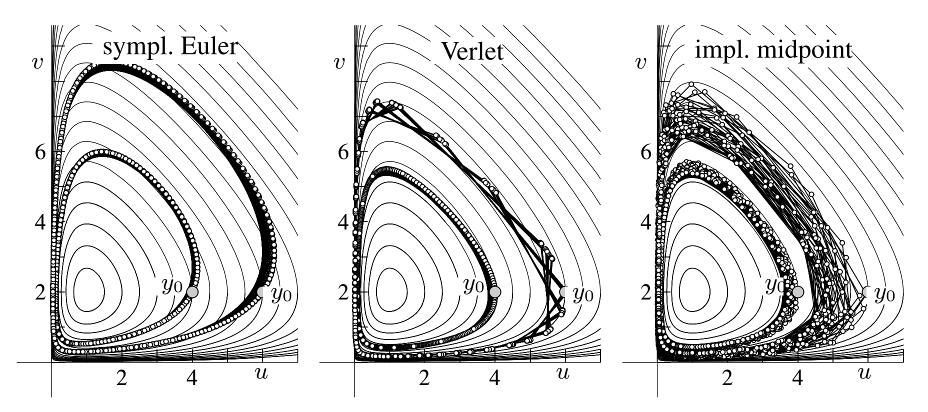


Fig. 4.1. Numerical solutions of the Lotka–Volterra equations (2.13) (step size $h = 0.25$, which is very large compared to the period of the solution; 1000 steps; initial values $(4, 2)$ and $(6, 2)$ for all methods)

$H^{[m]}(y)$, such that the individual systems $\dot{y} = B(y)\nabla H^{[i]}(y)$ can be solved exactly. The flow of these subsystems is a Poisson map and automatically respects the Casimirs, and so does their composition. McLachlan (1993), Reich (1993), and McLachlan & Quispel (2002) present several interesting examples.

Example 4.8. In the previous example of a Lotka–Volterra equation with separable Hamiltonian $H(u, v) = K(u) + L(v)$, the systems with Hamiltonian $K(u)$ and $L(v)$ can be solved explicitly. Since the flow of each of the subsystems is a Poisson map, so is their composition. Combining a half-step with L, a full step with K, and again a half-step with L, we thus obtain the following Verlet-like second-order Poisson integrator:

$$\begin{aligned} u_{n+1/2} &= \exp\bigl(\frac{h}{2}\, v_n \nabla L(v_n)\bigr) u_n \\ v_{n+1} &= \exp\bigl(-h u_{n+1/2} \nabla K(u_{n+1/2})\bigr) v_n \\ u_{n+1} &= \exp\bigl(\frac{h}{2}\, v_{n+1} \nabla L(v_{n+1})\bigr) u_{n+1/2}. \end{aligned} \tag{4.6}$$

In the setting of Hamiltonian systems on a manifold, the splitting approach can be formulated in the following way.

Variational Splitting. Consider a Hamiltonian system (2.17) on a symplectic manifold $\mathcal{M}$, and use a splitting $H = H^{[1]} + H^{[2]}$ of the Hamiltonian in the following algorithm:

1. Let $x_n^+ \in \mathcal{M}$ be the solution at time $h/2$ of the equation for x,

$$(J\dot{x} - \nabla H^{[1]}(x), \xi) = 0 \quad \text{for all } \xi \in T_x\mathcal{M} \tag{4.7}$$

   with initial value $x(0) = x_n$.

2. Let x_{n+1}^- be the solution at time h of
$$(J\dot{x} - \nabla H^{[2]}(x), \xi) = 0 \quad \text{for all } \xi \in T_x\mathcal{M} \tag{4.8}$$
with initial value $x(0) = x_n^+$.
3. Take x_{n+1} as the solution at time $h/2$ of (4.7) with initial value $x(0) = x_{n+1}^-$.

Splitting algorithms for Hamiltonian systems on manifolds have been studied by Dullweber, Leimkuhler & McLachlan (1997) and Benettin, Cherubini & Fassò (2001) in the context of rigid body dynamics; see Sect. VII.5. Lubich (2004) and Faou & Lubich (2004) have studied the above splitting method for applications in quantum molecular dynamics; see Sect. VII.6 for an example.

By Theorem 2.8, the substeps 1.–3. written in coordinates $x = \chi(y)$ are Poisson systems $\dot{y} = B(y)\nabla K^{[i]}(y)$ with $K^{[i]}(y) = H^{[i]}(\chi(y))$, but the algorithm itself is independent of the choice of coordinates. Since the substeps are exact flows of Hamiltonian systems on the manifold $\mathcal{M}$, their composition yields a symplectic map. In the coordinates y the substeps are the exact flows of Poisson systems, and hence their composition yields a Poisson map.

Poisson Integrators and Symplectic Integrators. Generally we note the following correspondence, which rephrases the remark on symplectic maps and Poisson maps after Definition 4.2. It applies in particular to the symplectic integrators for constrained mechanics of Sect. VII.1.

Lemma 4.9. *An integrator $x_1 = \Psi_h(x_0)$ for a Hamiltonian system (2.17) on a manifold $\mathcal{M}$ is symplectic if and only if the integrator written in local coordinates, $y_1 = \Phi_h(y_0)$ corresponding to a coordinate map $x = \chi(y)$, is a Poisson integrator for the structure matrix $B(y)$ of (2.21).*

VII.4.3 Integrators Based on the Darboux–Lie Theorem

If we explicitly know a transformation $z = \vartheta(y)$ that brings the system $\dot{y} = B(y)\nabla H(y)$ to canonical form (as in Corollary 3.6), we can proceed as follows: compute $z_n = \vartheta(y_n)$; apply a symplectic integrator to the transformed system $\dot{z} = B_0\nabla K(z)$ (B_0 is the matrix (3.18) and $K(z) = H(y)$) which yields $z_{n+1} = \Psi_h(z_n)$; compute finally y_{n+1} from $z_{n+1} = \vartheta(y_{n+1})$. This yields a Poisson integrator by the following lemma.

Lemma 4.10. *Let $z = (u, c) = \vartheta(y)$ be the transformation of Theorem 3.4. Suppose that the integrator $\Phi_h(y)$ takes the form*
$$\Psi_h(z) = \begin{pmatrix} \Psi_h^1(u, c) \\ c \end{pmatrix}$$
in the new variables $z = (u, c)$. Then, $\Phi_h(y)$ is a Poisson integrator if and only if $u \mapsto \Psi_h^1(u, c)$ is a symplectic integrator for every c.

Proof. The integrator $\Phi_h(y)$ is Poisson for the structure matrix $B(y)$ if and only if $\Psi_h(z)$ is Poisson for the matrix B_0 of (3.18); see Exercise 7. By assumption, $\Psi_h(z)$ preserves the Casimirs of B_0. The identity

$$\Psi_h'(z)B_0\Psi_h'(z)^T = \begin{pmatrix} A\,J^{-1}A^T & 0 \\ 0 & 0 \end{pmatrix}$$

with $A = \partial\Psi_h^1/\partial u$ proves the statement. □

Notice that the transformation ϑ has to be global in the sense that it has to be the same for all integration steps. Otherwise a degradation in performance, similar to that of the experiment in Example V.4.3, has to be expected.

Example 4.11. As a first illustration consider the Lotka–Volterra system (2.13). Applying the transformation $\vartheta(u,v) = (\ln u, \ln v) = (p,q)$, this system becomes canonically Hamiltonian with

$$K(p,q) = -H(u,v) = -H(e^p, e^q).$$

If we apply the symplectic Euler method to this Hamiltonian system, and if we transform back to the original variables, we obtain the method

$$\begin{aligned} u_{n+1} &= u_n \exp\big(hv_n H_v(u_{n+1}, v_n)\big), \\ v_{n+1} &= v_n \exp\big(-hu_{n+1} H_u(u_{n+1}, v_n)\big). \end{aligned} \tag{4.9}$$

In contrast to the method of Example 4.7, (4.9) is also a Poisson integrator for (2.13) if $H(u,v)$ is not separable. If we compose a step with step size $h/2$ of the symplectic Euler method with its adjoint method, then we obtain again, in the case of a separable Hamiltonian, the method (4.6).

Example 4.12 (Ablowitz–Ladik Discrete Nonlinear Schrödinger Equation). An interesting space discretization of the nonlinear Schrödinger equation is the Ablowitz–Ladik model

$$i\,\dot y_k + \frac{1}{\Delta x^2}\big(y_{k+1} - 2y_k + y_{k-1}\big) + |y_k|^2\big(y_{k+1} + y_{k-1}\big) = 0,$$

which we consider under periodic boundary conditions $y_{k+N} = y_k$ ($\Delta x = 1/N$). It is completely integrable (Ablowitz–Ladik 1976) and, as we shall see below, it is a Poisson system with noncanonical Poisson bracket. Splitting the variables into real and imaginary parts, $y_k = u_k + iv_k$, we obtain

$$\begin{aligned} \dot u_k &= -\frac{1}{\Delta x^2}\big(v_{k+1} - 2v_k + v_{k-1}\big) - \big(u_k^2 + v_k^2\big)\big(v_{k+1} + v_{k-1}\big) \\ \dot v_k &= \frac{1}{\Delta x^2}\big(u_{k+1} - 2u_k + u_{k-1}\big) + \big(u_k^2 + v_k^2\big)\big(u_{k+1} + u_{k-1}\big). \end{aligned}$$

With $u = (u_1, \dots, u_N)$, $v = (v_1, \dots, v_N)$ this system can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & -D(u,v) \\ D(u,v) & 0 \end{pmatrix} \begin{pmatrix} \nabla_u H(u,v) \\ \nabla_v H(u,v) \end{pmatrix}, \tag{4.10}$$

where $D = \mathrm{diag}(d_1, \ldots, d_N)$ is the diagonal matrix with entries

$$d_k(u,v) = 1 + \Delta x^2 \big(u_k^2 + v_k^2\big),$$

and the Hamiltonian is

$$H(u,v) = \frac{1}{\Delta x^2} \sum_{l=1}^{N} \Big(u_l u_{l-1} + v_l v_{l-1}\Big) - \frac{1}{\Delta x^4} \sum_{l=1}^{N} \ln\big(1 + \Delta x^2 (u_l^2 + v_l^2)\big).$$

We thus get a Poisson system (the conditions of Lemma 2.3 are directly verified). There are many possibilities to transform this system to canonical form. Tang, Pérez-García & Vázquez (1997) propose the transformation

$$p_k = \frac{1}{\Delta x \sqrt{1 + \Delta x^2 v_k^2}} \arctan\Big(\frac{\Delta x}{\sqrt{1 + \Delta x^2 v_k^2}} \cdot u_k\Big), \qquad q_k = v_k,$$

for which the inverse can be computed in a straightforward way. Here, we suggest the transformation

$$\begin{aligned} p_k &= u_k\, \sigma\big(\Delta x^2 (u_k^2 + v_k^2)\big) \\ q_k &= v_k\, \sigma\big(\Delta x^2 (u_k^2 + v_k^2)\big) \end{aligned} \qquad \text{with} \qquad \sigma(x) = \sqrt{\frac{\ln(1+x)}{x}}, \tag{4.11}$$

which treats the variables more symmetrically. Its inverse is

$$\begin{aligned} u_k &= p_k\, \tau\big(\Delta x^2 (p_k^2 + q_k^2)\big) \\ v_k &= q_k\, \tau\big(\Delta x^2 (p_k^2 + q_k^2)\big) \end{aligned} \qquad \text{with} \qquad \tau(x) = \frac{\exp x - 1}{x}.$$

Both transformations take the system (4.10) to canonical form. For the transformation (4.11) the Hamiltonian in the new variables is

$$\begin{aligned} H(p,q) = \frac{1}{\Delta x^2} \sum_{l=1}^{N} \tau\big(\Delta x^2 (p_l^2 + q_l^2)\big) \tau\big(\Delta x^2 (p_{l-1}^2 + q_{l-1}^2)\big) \Big(p_l p_{l-1} + q_l q_{l-1}\Big) \\ - \frac{1}{\Delta x^2} \sum_{l=1}^{N} \big(p_l^2 + q_l^2\big). \end{aligned}$$

Applying standard symplectic schemes to this Hamiltonian yields Poisson integrators for (4.10).

VII.5 Rigid Body Dynamics and Lie–Poisson Systems

> ... these topics, which, after all, have occupied workers in geometric mechanics for many years. (R. McLachlan 2003)

An important Poisson system is given by Euler's famous equations for the motion of a rigid body (see left picture of Fig. 5.1), for which we recall the history and derivation and present various structure-preserving integrators. Euler's equations are a particular case of Lie–Poisson systems, which result from a reduction process of Hamiltonian systems on a Lie group.

VII.5.1 History of the Euler Equations

> "Le sujet que je me propose de traiter ici, est de la derniere importance dans la Mécanique ; & j'ai déjà fait plusieurs efforts pour le mettre dans tout son jour. Mais, quoique le calcul ait assès bien réussi, & que j'ai découvert des formules analytiques ..., leur application étoit pourtant assujettie à des difficultés qui m'ont paru presque tout à fait insurmontables. Or, depuis que j'ai dévelopé les principes de la conoissance mécanique des corps, la belle propriété des trois axes principaux dont chaque corps est doué, m'a enfin mis en état de vaincre toutes ces difficultés, ..."
>
> (Euler 1758b, p. 154)

A great challenge for Euler were his efforts to establish a mathematical analysis for the motion of a rigid body. Due to the fact that such a body can have an arbitrary shape and mass distribution (see left picture of Fig. 5.2), and that the rotation axis can arbitrarily move with time, the problem is difficult and Euler struggled for many years (all these articles are collected in *Opera Omnia*, Ser. II, Vols. 7 and 8). The breakthrough was enabled by the discovery that any body, as complicated as may be its configuration, reduces to an inertia ellipsoid with three principal axes and three numbers, the principal moments of inertia (Euler 1758a; see the middle picture of Fig. 5.2 and the citation).

$$dx + \frac{cc-bb}{aa}.\, yz\,dt = \frac{2gP\,dt}{Maa}$$
$$dy + \frac{aa-cc}{bb}.\, xz\,dt = \frac{2gQ\,dt}{Mbb}$$
$$dz + \frac{bb-aa}{cc}.\, xy\,dt = \frac{2gR\,dt}{Mcc}$$

$\alpha p' = Ap' - Hq' - Gr'$;

, la quatrième et la si
, on aura pareillement

$\alpha q' = Bq' - Fr' - Hp'$;

me, la cinquième et la
$'$, q', on aura

$\alpha r' = Cr' - Gp' - Fq'$;

Fig. 5.1. Left picture: first publication of the Euler equations in Euler (1758b). Right picture: principal axes as eigenvectors in Lagrange (1788)

The Inertia Ellipsoid. We choose a moving coordinate system connected to the body $\mathcal{B}$ and we consider motions of the body where the origin is fixed. By another of Euler's famous theorems, any such motion is infinitesimally a rotation around an axis. We represent the rotation axis of the body by the *direction* of a vector ω and the speed of rotation by the *length* of ω. Then the velocity of a mass point x of $\mathcal{B}$ is given by the exterior product

$$v = \omega \times x = \begin{pmatrix} \omega_2 x_3 - \omega_3 x_2 \\ \omega_3 x_1 - \omega_1 x_3 \\ \omega_1 x_2 - \omega_2 x_1 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{5.1}$$

(orthogonal to ω, orthogonal to x, and of length $\|\omega\| \cdot \|x\| \cdot \sin\gamma$; see the left picture of Fig. 5.2). The *kinetic energy* is obtained by integrating the energy of the mass

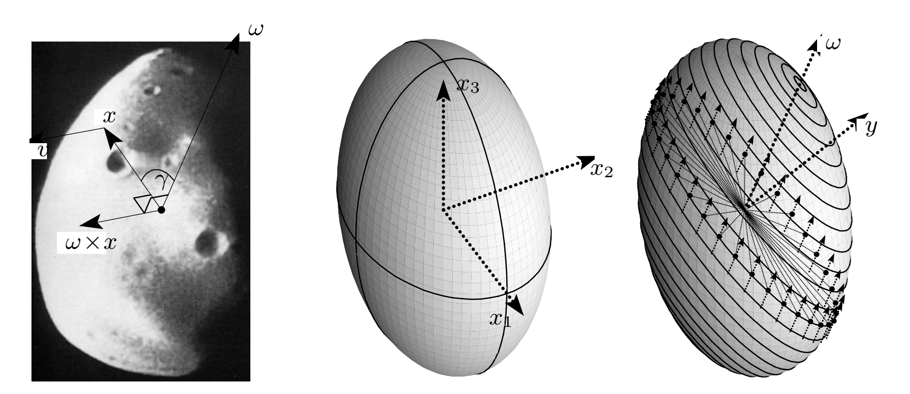


Fig. 5.2. A rigid body rotating around a variable axis (left); the corresponding inertia ellipsoid (middle); the corresponding angular momentum (right)

points dm over the body

$$\begin{aligned} T &= \frac{1}{2}\int_{\mathcal{B}} \|\omega \times x\|^2 \, dm \qquad (5.2) \\ &= \frac{1}{2}\int_{\mathcal{B}} \Big((\omega_2 x_3 - \omega_3 x_2)^2 + (\omega_3 x_1 - \omega_1 x_3)^2 + (\omega_1 x_2 - \omega_2 x_1)^2 \Big) \, dm \, . \end{aligned}$$

If this is multiplied out, one obtains

$$T = \frac{1}{2}\,\omega^T \Theta \omega\,, \text{ where } \Theta_{ii} = \int_{\mathcal{B}} (x_k^2 + x_\ell^2)\, dm,\ \Theta_{ik} = -\int_{\mathcal{B}} x_i x_k \, dm, \ \ (i \neq k, \ell). \qquad (5.3)$$

Euler (1758a) showed, by endless trigonometric transformations, that there exist principal axes of the body in which this expression takes the form

$$T = \frac{1}{2}\Big(I_1 \omega_1^2 + I_2 \omega_2^2 + I_3 \omega_3^2 \Big)\,. \qquad (5.4)$$

This was historically the first transformation of such a 3×3 quadratic form to diagonal form. Later, Lagrange (1788) discovered that these axes were the eigenvectors of the matrix Θ and the moments of inertia I_k the corresponding eigenvalues (without calling them so, see the right picture of Fig. 5.1).

The Angular Momentum. The first law of Newton's *Principia* states that the *momentum* $v \cdot m$ of a mass point remains constant in the absence of exterior forces. The corresponding quantity for *rotational* motion is the *angular momentum*, i.e., the exterior product $x \times v$ times the mass. Integrating over the body we obtain, with (5.1),

$$y = \int_{\mathcal{B}} (x \times v) \, dm = \int_{\mathcal{B}} \Big(x \times (\omega \times x) \Big) \, dm\,. \qquad (5.5)$$

If this is multiplied out, the matrix Θ appears again and one obtains the surprising result (due to Poinsot 1834)

$$y = \Theta\,\omega, \qquad \text{or, in the principal axes coordinates,} \qquad y_k = I_k\,\omega_k\,. \tag{5.6}$$

Such a relation is familiar from the theory of conjugate diameters (Apollonius, Book II, Prop. VI): the angular momentum is a vector orthogonal to the plane of vectors conjugate to ω (see the right picture of Fig. 5.2).

The Euler Equations. Euler's paper (1758a), on his discovery of the principal axes, is immediately followed by Euler (1758b), where he derives his equations for the motion of a rigid body by long, doubtful and often criticized calculations, repeated in a little less doubtful manner in Euler's monumental treatise (1765). Beauty and elegance, not only of the result, but also of the proof, is due to Poinsot (1834) and Hayward (1856). It is masterly described by Klein & Sommerfeld (1897), and in Chapter 6 of Arnold (1989).

From now on we choose the coordinate system, moving with the body, such that the inertia tensor remains diagonal. We also watch the motion of the body from a coordinate system stationary in the space. The transformation of a vector $x \in \mathbb{R}^3$ in the body frame [4], to the corresponding $\widetilde{x} \in \mathbb{R}^3$ in the stationary frame, is denoted by

$$\widetilde{x} = Q(t)x\,. \tag{5.7}$$

The matrix $Q(t)$ is orthogonal and describes the motion of the body: for $x = e_i$ we see that the columns of $Q(t)$ are the coordinates of the body's principal axes in the stationary frame.

The analogous statement to Newton's first law for rotational motion is: *in the absence of exterior angular forces, the angular momentum $\widetilde{y}$, seen from the fixed coordinate system, is a constant vector* [5]. This same vector y, seen from the *moving* frame, which at any instance rotates with the body around the vector ω, rotates in the *opposite* direction. Therefore we have from (5.1), by changing the signs of ω, the derivatives

$$\begin{pmatrix} \dot y_1 \\ \dot y_2 \\ \dot y_3 \end{pmatrix} = \begin{pmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \tag{5.8}$$

If we insert $\omega_k = y_k/I_k$ from (5.6), we obtain

$$\begin{pmatrix} \dot y_1 \\ \dot y_2 \\ \dot y_3 \end{pmatrix} = \begin{pmatrix} 0 & y_3/I_3 & -y_2/I_2 \\ -y_3/I_3 & 0 & y_1/I_1 \\ y_2/I_2 & -y_1/I_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} (I_3^{-1} - I_2^{-1})\, y_3 y_2 \\ (I_1^{-1} - I_3^{-1})\, y_1 y_3 \\ (I_2^{-1} - I_1^{-1})\, y_2 y_1 \end{pmatrix} \tag{5.9}$$

or, by rearranging the products the other way round,

$$\begin{pmatrix} \dot y_1 \\ \dot y_2 \\ \dot y_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1/I_1 \\ y_2/I_2 \\ y_3/I_3 \end{pmatrix}, \tag{5.10}$$

[4] Long-standing tradition, from Klein to Arnold, uses capitals for denoting the coordinates in this moving frame; but this would lead to confusion with our subsequent matrix notation

[5] For a proof of this statement by d'Alembert's Principle, see Sommerfeld (1942), §II.13.

written in two different ways as a Poisson system, whose right hand vectors are the gradients of $C(y) = \frac{1}{2}\sum_{k=1}^{3} y_k^2$ and $H(y) = \frac{1}{2}\sum_{k=1}^{3} I_k^{-1} y_k^2$, respectively. These are the two quadratic invariants of Chap. IV. The first represents the length of the constant angular momentum $\widetilde{y}$ in the orthogonal body frame, and the second represents the energy (5.4).

Computation of the Position Matrix $Q(t)$**.** Once we have solved the Euler equations for $y(t)$, we obtain the rotation vector $\omega(t)$ by (5.6). It remains to find the matrix $Q(t)$ which gives the position of our rotating body. We know that the columns of the matrix Q, seen in the stationary frame, correspond to the unit vectors e_i in the body frame. These rotate, by (5.1), with the velocity

$$\Big(\omega\times e_1, \omega\times e_2, \omega\times e_3\Big) = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} =: W\,. \tag{5.11}$$

We thus obtain $\dot{Q}$, the rotational velocity expressed in the stationary frame, by the back transformation (5.7):

$$\dot{Q} = QW \qquad \text{or} \qquad Q^T\dot{Q} = W\,. \tag{5.12}$$

This is a differential system for Q which, because W is skew-symmetric, preserves the orthogonality of Q. The problem is solved – in theory.

VII.5.2 Hamiltonian Formulation of Rigid Body Motion

In order to open the door for efficient numerical algorithms, we treat the rigid body as a constrained Hamiltonian system.

Position Variables. The position of the rigid body at time t is determined, in view of (5.7), by a three-dimensional orthogonal matrix $Q(t)$. The constraints to be respected are thus $Q^TQ - I = 0$.

Kinetic Energy. As in (5.12), we associate with Q and $\dot{Q}$ the skew-symmetric matrix $W = Q^T\dot{Q}$ whose entries ω_k, arranged as in (5.11), determine the kinetic energy by (5.4):

$$T = \frac{1}{2}\big(I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2\big)\,.$$

For any diagonal matrix $D = \mathrm{diag}(d_1, d_2, d_3)$ we observe

$$\mathrm{trace}\,(WDW^T) = (d_2+d_3)\omega_1^2 + (d_3+d_1)\omega_2^2 + (d_1+d_2)\omega_3^2\,.$$

Therefore, with

$$I_1 = d_2+d_3, \quad I_2 = d_3+d_1, \quad I_3 = d_1+d_2 \quad \text{or} \quad d_k = \int_{\mathcal{B}} x_k^2\,dm \tag{5.13}$$

(note that $d_k > 0$ for all bodies that have interior points), we obtain the kinetic energy as

$$T = \frac{1}{2}\,\text{trace}\,(WDW^T)\,. \tag{5.14}$$

Inserting $W = Q^T\dot{Q}$, we have

$$T = \frac{1}{2}\,\text{trace}\,(Q^T\dot{Q}D\dot{Q}^TQ) = \frac{1}{2}\,\text{trace}\,(\dot{Q}D\dot{Q}^T)\,, \tag{5.15}$$

since Q is an orthogonal matrix.

Conjugate Variables. We now have an expression for the kinetic energy in terms of derivatives of position coordinates and are able to introduce the conjugate momenta

$$P = \partial T/\partial\dot{Q} = \dot{Q}D\,. \tag{5.16}$$

If we suppose to have, in addition to T, a potential $U(Q)$, we get the Hamiltonian

$$H(P,Q) = \frac{1}{2}\,\text{trace}\,\big(PD^{-1}P^T\big) + U(Q)\,. \tag{5.17}$$

Lagrange Multipliers. The constraints are given by the orthogonality of Q, i.e., the equation $g(Q) = Q^TQ - I = 0$. Since this matrix is always symmetric, this consists of $\frac{1}{2}n(n+1) = 6$ independent algebraic conditions, calling for six Lagrange multipliers. If the expression $G(Q)^T\lambda$ in (1.9) is actually computed, it turns out that this term becomes the product $Q\Lambda$, where the six Lagrange multipliers are arranged in a symmetric matrix Λ; see also formula (IV.9.6). Thus, the constrained Hamiltonian system (1.9) reads in our case, with $\nabla U = \big(\partial U/\partial Q_{ij}\big)$,

$$\begin{aligned} \dot{Q} &= PD^{-1} \\ \dot{P} &= -\nabla U(Q) - Q\Lambda \qquad (\Lambda \text{ symmetric}) \\ 0 &= Q^TQ - I\,. \end{aligned} \tag{5.18}$$

Reduction to the Euler Equations. The key idea is to introduce the matrix

$$Y = Q^TP = Q^T\dot{Q}D = WD = \begin{pmatrix} 0 & -d_2\omega_3 & d_3\omega_2 \\ d_1\omega_3 & 0 & -d_3\omega_1 \\ -d_1\omega_2 & d_2\omega_1 & 0 \end{pmatrix}\,, \tag{5.19}$$

where the ω_k can be further expressed in terms of the angular momenta $y_k = I_k\,\omega_k$. Using the notation $\text{skew}\,(A) = \frac{1}{2}(A - A^T)$, we see, with (5.13), that

$$Y - Y^T = 2\,\text{skew}\,(Y) = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \tag{5.20}$$

contains just the angular momenta. Moreover, DY is skew-symmetric. By (5.18) the derivative of Y is seen to be

$$\dot{Y} = \dot{Q}^T P + Q^T \dot{P} = D^{-1} P^T P - Q^T \nabla U(Q) - \Lambda = D^{-1} Y^T Y - Q^T \nabla U(Q) - \Lambda .$$

Taking the skew-symmetric part of this equation, the symmetric matrix Λ drops out and we obtain

$$\text{skew}\,(\dot{Y}) = \text{skew}\,(D^{-1} Y^T Y) - \text{skew}\,(Q^T \nabla U(Q)) . \qquad (5.21)$$

These are, for $U = 0$, precisely the above Euler equations, obtained a second time.

VII.5.3 Rigid Body Integrators

For a numerical simulation of rigid body motions, one can either solve the constrained Hamiltonian system (5.18), or one can solve the differential equation (5.21) for the angular momentum $Y(t)$ in tandem with the equation (5.12) for $Q(t)$. We consider the following approaches: (I) an efficient application of the RATTLE algorithm (1.26), and (II) various splitting methods.

(I) RATTLE

We apply the symplectic RATTLE algorithm (1.26) to the system (5.18), and rewrite the formulas in terms of the variables Y and Q. This approach has been proposed and developed independently by McLachlan & Scovel (1995) and Reich (1994).

An application of the RATTLE algorithm (1.26) to the system (5.18) yields

$$\begin{aligned} P_{1/2} &= P_0 - \frac{h}{2} \nabla U(Q_0) - \frac{h}{2} Q_0 \Lambda_0 \\ Q_1 &= Q_0 + h P_{1/2} D^{-1}, \qquad Q_1^T Q_1 = I \\ P_1 &= P_{1/2} - \frac{h}{2} \nabla U(Q_1) - \frac{h}{2} Q_1 \Lambda_1, \qquad Q_1^T P_1 D^{-1} + D^{-1} P_1^T Q_1 = 0, \end{aligned} \qquad (5.22)$$

where both Λ_0 and Λ_1 are symmetric matrices. We let $Y_0 = Q_0^T P_0$, $Y_1 = Q_1^T P_1$, and $Z = Q_0^T P_{1/2} D^{-1}$. We multiply the first relation of (5.22) by Q_0^T, the last relation by Q_1^T, and we eliminate the symmetric matrices Λ_0 and Λ_1 by taking the skew-symmetric parts of the resulting equations. The orthogonality of $Q_0^T Q_1 = I + hZ$ implies $hZ^T Z = -(Z + Z^T)$, which can then be used to simplify the last relation. Altogether this results in the following algorithm.

Algorithm 5.1. *Let Q_0 be orthogonal and DY_0 be skew-symmetric. One step $(Q_0, Y_0) \mapsto (Q_1, Y_1)$ of the method then reads as follows:*

– *find Z such that $I + hZ$ is orthogonal and*

$$\text{skew}\,(ZD) = \text{skew}\,(Y_0) - \frac{h}{2}\,\text{skew}\,\big(Q_0^T \nabla U(Q_0)\big), \qquad (5.23)$$

– *compute $Q_1 = Q_0(I + hZ)$,*
– *compute Y_1 such that DY_1 is skew-symmetric and*

$$\text{skew}\,(Y_1) = \text{skew}\,(ZD) - \text{skew}\,\big((Z + Z^T)D\big) - \frac{h}{2}\,\text{skew}\,\big(Q_1^T \nabla U(Q_1)\big).$$

The second step is explicit, and the third step represents a linear equation for the elements of Y_1.

Computation of the First Step. We write for the known part of equation (5.23)

$$\text{skew}\,(Y_0) - \frac{h}{2}\,\text{skew}\,\big(Q_0^T \nabla U(Q_0)\big) \;=\; \begin{pmatrix} 0 & -\alpha_3 & \alpha_2 \\ \alpha_3 & 0 & -\alpha_1 \\ -\alpha_2 & \alpha_1 & 0 \end{pmatrix} = A \qquad (5.24)$$

and have to solve

$$\frac{1}{2}(ZD - DZ^T) = A\,, \quad (I + hZ^T)(I + hZ) = I\,, \quad \frac{1}{2}(ZD + DZ^T) = S$$

(the trick was to add the last equation with S an unknown symmetric matrix). Elimination gives $Z = (A + S)D^{-1}$ and $Z^T = D^{-1}(S - A)$. Both inserted into the second equation lead to a Riccati equation for S. There exist efficient algorithms for such problems; see the reference in Sect. IV.5.3 and a detailed explanation in McLachlan & Zanna (2005).

Remark 5.2 (Moser–Veselov Algorithm). An independent access to the above formulas is given in a remarkable paper by Moser & Veselov (1991), by treating the rigid body through a *discretized* variational principle, similar to the ideas of Sect. VI.6.2. The equivalence is explained by McLachlan & Zanna (2005), following a suggestion of B. Leimkuhler and S. Reich.

Quaternions (Euler Parameters). An efficient implementation of the above algorithm requires suitable representations of orthogonal matrices, and the use of quaternions is a standard approach.

After having revolutionized Lagrangian mechanics (see Chapt. VI), Hamilton struggled for years to generalize complex analysis to three dimensions. He finally achieved his dream, however not in three dimensions, but in *four*, and founded in 1843 the theory of quaternions.

For an introduction to quaternions (whose coefficients are sometimes called Euler parameters) we refer to Sects. IV.2 and IV.3 of Klein (1908), and for their use in numerical simulations to Sects. 9.3 and 11.3 of Haug (1989). Quaternions can be written as $e = e_0 + ie_1 + je_2 + ke_3$, where multiplication is defined via the relations $i^2 = j^2 = k^2 = -1$, $ij = k$, $jk = i$, $ki = j$, and $ji = -k$, $kj = -i$, $ik = -j$. The product of two quaternions $e \cdot f$ (written in matrix notation) is

$$\begin{array}{l}(e_0 + ie_1 + je_2 + ke_3) \\ \quad\cdot(f_0 + if_1 + jf_2 + kf_3)\end{array} \;=\; \begin{pmatrix} e_0 & -e_1 & -e_2 & -e_3 \\ e_1 & e_0 & -e_3 & e_2 \\ e_2 & e_3 & e_0 & -e_1 \\ e_3 & -e_2 & e_1 & e_0 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix} \qquad (5.25)$$

We see (in grey) that in dimensions $1, 2, 3$ appears a skew-symmetric matrix E whose structure is familiar to us. This part of the matrix changes sign if the two factors are permuted.

An important discovery, for three dimensional applications of the quaternions, is the following: if a quaternion p is a 3-vector (i.e., has $p_0 = 0$), then $p' = e\cdot p\cdot \overline{e}$ is a 3-vector, too, and the map $p \mapsto p'$ is described by the matrix

$$Q(e) \,=\, \|e\|^2 I + 2e_0 E + 2E^2, \qquad E \,=\, \begin{pmatrix} 0 & -e_3 & e_2 \\ e_3 & 0 & -e_1 \\ -e_2 & e_1 & 0 \end{pmatrix} \tag{5.26}$$

where $\overline{e} = e_0 - ie_1 - je_2 - ke_3$ and $\|e\|^2 = e\cdot\overline{e} = e_0^2 + e_1^2 + e_2^2 + e_3^2$.

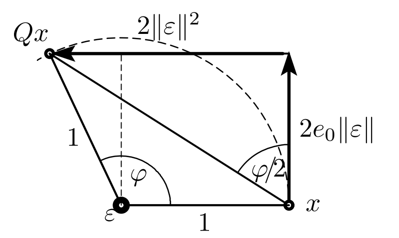
Lemma 5.3. *If $\|e\| = 1$, then the matrix $Q(e)$ is orthogonal. Every orthogonal matrix with $\det Q = 1$ can be written in this form. We have $Q(e)Q(f) = Q(ef)$, so that the multiplication of orthogonal matrices corresponds to the multiplication of quaternions.*

Geometrically, the matrix Q effects a rotation around the axis $\varepsilon = (e_1, e_2, e_3)^T$ with rotation angle φ which satisfies $\tan(\varphi/2) = \|\varepsilon\|/e_0$.

Proof. The condition $Q^T Q = I$ can be verified directly using $E^T = -E$ and $E^3 = -(e_1^2 + e_2^2 + e_3^2)E$. The reciprocal statement is a famous theorem of Euler; it is based on the fact that ε is an eigenvector of Q, which in dimension 3×3 always exists. The formula for $Q(e)Q(f)$ follows from $e\cdot f\cdot p\cdot\overline{f}\cdot\overline{e} = (e\cdot f)\cdot p\cdot(\overline{e\cdot f})$.

The geometric property follows from the virtues of the exterior product, because by (5.1) the matrix Q maps a vector x to

$$x + 2e_0\; \varepsilon\times x + 2\;\varepsilon\times(\varepsilon\times x).$$



This consists in a rectangular mouvement in a plane orthogonal to ε; first vertical to x by an amount $2e_0\,\|\varepsilon\|$ (times the distance of x), then parallel to x by an amount $2\|\varepsilon\|^2$.

Applying Pythagoras' Theorem as $(2e_0\,\|\varepsilon\|)^2 + (2\|\varepsilon\|^2 - 1)^2 = 1$, it turns out that the map is norm preserving if $e_0^2 + \|\varepsilon\|^2 = 1$. The angle $\varphi/2$, whose tangens can be seen to be $\|\varepsilon\|/e_0$, is an angle at the circumference of the circle for the rotation angle φ at the center. □

For an efficient implementation of Algorithm 5.1 we represent the orthogonal matrices Q_0, Q_1, and $I + hZ$ by quaternions. This reduces the dimension of the systems, and step 2 becomes a simple multiplication of quaternions. For solving the nonlinear system of step 1, we let $I + hZ = Q(e)$. With the values of α_i from (5.24) and with $\operatorname{skew}(hZD) = 2e_0 \operatorname{skew}(ED) + 2\operatorname{skew}(E^2 D)$, the equation (5.23) becomes

$$\begin{pmatrix} h\alpha_1 \\ h\alpha_2 \\ h\alpha_3 \end{pmatrix} = 2e_0 \begin{pmatrix} I_1 e_1 \\ I_2 e_2 \\ I_3 e_3 \end{pmatrix} + 2 \begin{pmatrix} (I_3 - I_2)e_2 e_3 \\ (I_1 - I_3)e_3 e_1 \\ (I_2 - I_1)e_1 e_2 \end{pmatrix}, \tag{5.27}$$

which, together with $e_0^2 + e_1^2 + e_2^2 + e_3^2 = 1$, represent four quadratic equations for four unknowns. We solve them very quickly by a few fixed-point iterations: update

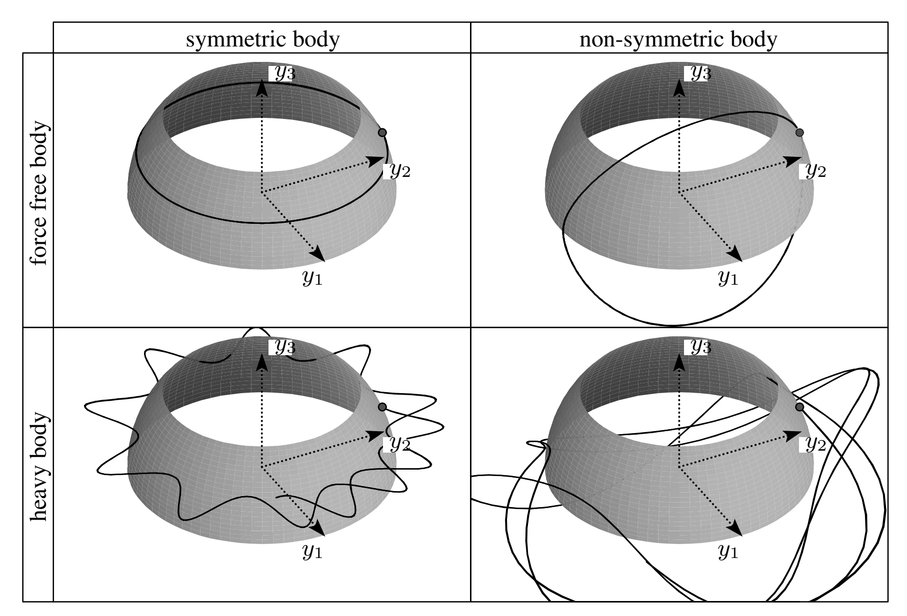


Fig. 5.3. Numerical solutions of the rigid body equations; without/with gravitation, with/without symmetry. Initial values $y_{10} = 0.2\,, y_{20} = 1.0\,, y_{30} = 0.4\,$; initial position of Q_0 determined by the quatenion $e_0 = 0.4\,, e_1 = 0.2\,, e_2 = 0.4\,, e_3 = 0.8\,$; moments of inertia $I_1 = 0.5$, $I_2 = 0.85$ (0.5 in the symmetric case), $I_3 = 1\,$; step size $h = 0.1$, integration interval $0 \le t \le 30$

successively e_i from the ith equation of (5.27) and then e_0 from the normalization condition. A Fortran subroutine RATORI for this algorithm is available on the homepage <http://www.unige.ch/∼hairer>.

Conservation of Casimir and Hamiltonian. It is interesing to note that, in the absence of a potential, the Algorithm 5.1 preserves exactly the Casimir $y_1^2 + y_2^2 + y_3^2$ and, more surprisingly, also the Hamiltonian $\frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$. This can be seen as follows: without any potential we have $\text{skew}\,(Y_0) = \text{skew}\,(ZD)$ and $\text{skew}\,(Y_1) = -\,\text{skew}\,(Z^T D)$, so that the vectors $(y_{10}, y_{20}, y_{30})^T$ and $(y_{11}, y_{21}, y_{31})^T$ are equal to $u+v$ and $u-v$, respectively, where u and v are the vectors of the right-hand side of (5.27). Since u and v are orthogonal, we have $\|u + v\| = \|u - v\|$, which proves the conservation of the Casimir.

To prove the conservation of the Hamiltonian, we first multiply the relation (5.27) with $G = \text{diag}(1/\sqrt{I_1}, 1/\sqrt{I_2}, 1/\sqrt{I_3})$, and then apply the same arguments. The vectors Gu and Gv are still orthogonal.

Example 5.4 (Force-Free and Heavy Top). We present in Fig. 5.3 the numerical solutions y_i obtained by the above algorithm. In the case of the heavy top, we assume the centre of gravity to be $(0, 0, 1)$ in the body frame, and assume that the third coordinate of the stationary frame is vertical. The potential energy due to gravity is

then given by $U(Q) = q_{33}$ and, expressed by quaternions (5.26), it is $U = e_0^2 - e_1^2 - e_2^2 + e_3^2$.

(II) Splitting Methods

As before we consider the differential equation (5.21) for the angular momenta in the body y_1, y_2, y_3 together with the differential equation (5.12) for the rotation matrix Q. An obvious splitting in the presence of a potential is

$$\varphi_{h/2}^U \circ \Phi_h^T \circ \varphi_{h/2}^U, \tag{5.28}$$

where φ_t^U represents the exact flow of

$$\dot{Q} = 0, \qquad \text{skew}\,(\dot{Y}) = -\,\text{skew}\,(Q^T \nabla U(Q)),$$

and Φ_h^T is a suitable numerical approximation of the system corresponding to kinetic energy only, i.e., without any potential $U(Q)$. The use of splitting techniques for rigid body dynamics was proposed by Touma & Wisdom (1994), McLachlan (1993), Reich (1994), and Dullweber, Leimkuhler & McLachlan (1997). Fassò (2003) presents a careful study and comparison of various ways of splitting the kinetic energy.

Computation of Φ_h^T. We do this by splitting once again, by letting several moments of inertia tending to infinity (and the corresponding ω_i tending to zero). In order to avoid formal difficulties with infinite denominators, we write the system (5.10) together with (5.12) in the form

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} T_{y_1}(y) \\ T_{y_2}(y) \\ T_{y_3}(y) \end{pmatrix} \tag{5.29}$$

$$\dot{Q} = Q \begin{pmatrix} 0 & -T_{y_3}(y) & T_{y_2}(y) \\ T_{y_3}(y) & 0 & -T_{y_1}(y) \\ -T_{y_2}(y) & T_{y_1}(y) & 0 \end{pmatrix}, \tag{5.30}$$

where $T(y) = \frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$ is the kinetic energy, and $T_{y_i}(y)$ denote partial derivatives.

Three Rotations Splitting. An obvious splitting of the kinetic energy is

$$T(y) = R_1(y) + R_2(y) + R_3(y), \qquad R_i(y) = y_i^2/(2I_i), \tag{5.31}$$

which results in the numerical method

$$\Phi_h^T = \varphi_{h/2}^{R_3} \circ \varphi_{h/2}^{R_2} \circ \varphi_h^{R_1} \circ \varphi_{h/2}^{R_2} \circ \varphi_{h/2}^{R_3},$$

where $\varphi_t^{R_i}$ is the exact flow of (5.29)-(5.30) with $T(y)$ replaced by $R_i(y)$. The flow $\varphi_t^{R_1}$ is easily obtained: y_1 remains constant and the second and third equation in (5.29) boil down to the harmonic oscillator. We obtain

$$y(t) = S(\alpha t)y(0), \qquad Q(t) = Q(0)S(\alpha t)^T \tag{5.32}$$

with $\alpha = y_1(0)/I_1$ and the rotation matrix

$$S(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}.$$

Similar simple formulas are obtained for the exact flows corresponding to R_2 and R_3.

Symmetric + Rotation Splitting. It is often advantageous, in particular for a nearly symmetric body ($I_1 \approx I_2$), to consider the splitting

$$T(y) = R(y) + S(y), \quad R(y) = \Big(\frac{1}{I_1} - \frac{1}{I_2}\Big)\frac{y_1^2}{2}, \quad S(y) = \frac{1}{2}\Big(\frac{y_1^2 + y_2^2}{I_2} + \frac{y_3^2}{I_3}\Big)$$

and the corresponding numerical integrator

$$\Phi_h^T = \varphi_{h/2}^R \circ \varphi_h^S \circ \varphi_{h/2}^R.$$

The exact flow φ_t^R is the same as (5.32) with I_1^{-1} replaced by $I_1^{-1} - I_2^{-1}$. The flow of the symmetric force-free top φ_t^S possesses simple analytic formulas, too (see the first picture of Fig. 5.3): we observe a precession of y with constant speed around a cone and a rotation of the body around ω with constant speed. Therefore

$$y(t) = B(\beta t)y(0), \qquad Q(t) = Q(0)A(t)B(\beta t)^T, \tag{5.33}$$

where $\beta = (I_3^{-1} - I_2^{-1})y_3(0)$, and

$$A(t) = \exp\left(\frac{t}{I_2}\begin{pmatrix} 0 & -y_3(0) & y_2(0) \\ y_3(0) & 0 & -y_1(0) \\ -y_2(0) & y_1(0) & 0 \end{pmatrix}\right), \quad B(\theta) = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This can also be checked directly by differentiation.

Similar to the previous algorithm it is advantageous to use a representation of the appearing orthogonal matrices by quaternions. The correspondence between the orthogonal rotation matrices appearing in (5.32) and (5.33) and their quaternions is, in accordance with Lemma 5.3, the following:

$$\begin{aligned} S(\theta)^T \quad &\leftrightarrow \quad \cos(\theta/2) + i\sin(\theta/2) \\ B(\theta)^T \quad &\leftrightarrow \quad \cos(\theta/2) + k\sin(\theta/2) \\ A(t) \quad &\leftrightarrow \quad \cos(\vartheta/2) + a^{-1}\sin(\vartheta/2)\big(iy_1(0) + jy_2(0) + ky_3(0)\big), \end{aligned}$$

where $a = \sqrt{y_1(0)^2 + y_2(0)^2 + y_3(0)^2}$ and $\vartheta = at/I_2$. The matrix multiplications in the algorithm can therefore be done very efficiently. A Fortran subroutine QUATER for the "symmetric + rotation splitting" algorithm is available on the homepage <http://www.unige.ch/∼hairer>.

VII.5.4 Lie–Poisson Systems

In Sect. VII.5.1 we have seen that the reduction of the equations of motion of the rigid body leads to the Poisson system (5.10) with a structure matrix whose entries are linear functions. Here we consider more general Poisson systems

$$\dot{y} = B(y)\nabla H(y), \tag{5.34}$$

where the structure matrix $B(y)$ depends linearly on y, i.e.,

$$b_{ij}(y) = \sum_{k=1}^{n} C_{ji}^{k}\, y_k \qquad \text{for } \ i,j = 1,\ldots,n. \tag{5.35}$$

Such systems, called Lie–Poisson systems, are closely related to differential equations on the dual of Lie algebras; see Marsden & Ratiu (1999), Chaps. 13 and 14, for an in-depth discussion of this theory.

Recall that a Lie algebra is a vector space with a bracket which is anti-symmetric and satisfies the Jacobi identity (Sect. IV.6). Let $E_1,\ldots,E_n$ be a basis of a vector space, and define a bracket by

$$[E_i,E_j] = \sum_{k=1}^{n} C_{ij}^{k} E_k \tag{5.36}$$

with C_{ij}^k from (5.35). If the structure matrix $B(y)$ of (5.35) is skew-symmetric and satisfies (2.10), then this bracket makes the vector space a Lie algebra (the verification is left as an exercise). The coefficients C_{ij}^k are called *structure constants* of the Lie algebra. Conversely, if we start from a Lie algebra with bracket given by (5.36), then the matrix $B(y)$ defined by (5.35) is the structure matrix of a Poisson bracket.

Let $\mathfrak{g}$ be a Lie algebra with a basis $E_1,\ldots,E_n$, and let $\mathfrak{g}^*$ be the dual of the Lie algebra, i.e., the vector space of all linear forms $Y:\mathfrak{g}\to\mathbb{R}$. The duality is written as $\langle Y,X\rangle$ for $Y\in\mathfrak{g}^*$ and $X\in\mathfrak{g}$. We denote by $F_1,\ldots,F_n$ the dual basis defined by $\langle F_i,E_j\rangle = \delta_{ij}$, the Kronecker δ.

Theorem 5.5. *Let $\mathfrak{g}$ be a Lie algebra with basis $E_1,\ldots,E_n$ satisfying (5.36). To $y=(y_1,\ldots,y_n)^T\in\mathbb{R}^n$ we associate $Y=\sum_{j=1}^n y_jF_j\in\mathfrak{g}^*$, and we consider a Hamiltonian*[6] *$H(y)=H(Y)$.*

Then, the Poisson system $\dot{y}=B(y)\nabla H(y)$ with $B(y)$ given by (5.35) is equivalent to the following differential equation on the dual $\mathfrak{g}^*$:

$$\langle \dot{Y},X\rangle = \langle Y,[H'(Y),X]\rangle \qquad \textit{for all } \ X\in\mathfrak{g}, \tag{5.37}$$

where $H'(Y)=\sum_{j=1}^n \frac{\partial H(y)}{\partial y_j}E_j$.

[6] We use the same symbol H for the functions $H:\mathbb{R}^n\to\mathbb{R}$ and $H:\mathfrak{g}^*\to\mathbb{R}$.

Proof. Differentiating $H(y) = H(Y)$ with respect to y_i gives

$$\frac{\partial H(y)}{\partial y_i} = \langle F_i, H'(Y)\rangle \qquad \text{and} \qquad H'(Y) = \sum_{j=1}^{n} \frac{\partial H(y)}{\partial y_j}\, E_j.$$

Here we have used the identification $(\mathfrak{g}^*)^* = \mathfrak{g}$, because $H'(Y)$ is actually an element of $(\mathfrak{g}^*)^*$. With this formula for $H'(Y)$ we are able to compute

$$\big\langle Y, [H'(Y), E_i]\big\rangle = \Big\langle Y, \sum_{j=1}^{n} \frac{\partial H(y)}{\partial y_j}\,[E_j, E_i]\Big\rangle = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial H(y)}{\partial y_j}\, C_{ji}^k\, \langle Y, E_k\rangle,$$

where we have used (5.36). Since $\langle \dot Y, E_i\rangle = \dot y_i$ and $\langle Y, E_k\rangle = y_k$, this shows that the differential equation (5.37) is equivalent to

$$\dot y_i = \sum_{j=1}^{n}\Big(\sum_{k=1}^{n} C_{ji}^k\, y_k\Big)\frac{\partial H(y)}{\partial y_j},$$

which is nothing more than $\dot y = B(y)\nabla H(y)$ with $B(y)$ from (5.35). □

We remark that (5.37) can be reformulated as

$$\dot Y = \mathrm{ad}^*_{H'(Y)} Y,$$

where ad^*_A is the adjoint of the operator $\mathrm{ad}_A(X) = [A, X]$.

Equation (5.37) is similar in appearance to the Lie bracket equation $\dot L = [A(L), L] = \mathrm{ad}_{A(L)} L$ of Sect. IV.3.2. When $\mathfrak{g}$ is the Lie algebra of a matrix Lie group G, then solutions of that equation are of the form $L(t) = \mathrm{Ad}_{U(t)} L_0$ where

$$\mathrm{Ad}_U X = U X U^{-1}; \tag{5.38}$$

see the proof of Lemma IV.3.4. Similarly, for the solution of (5.37) we have the following.

Theorem 5.6. *Consider a matrix Lie group G with Lie algebra $\mathfrak{g}$. Then, the solution $Y(t) \in \mathfrak{g}^*$ of (5.37) with initial value $Y_0 \in \mathfrak{g}^*$ is given by*

$$\langle Y(t), X\rangle = \langle Y_0, U(t)^{-1} X U(t)\rangle \qquad \textit{for all } X \in \mathfrak{g}, \tag{5.39}$$

where $U(t) \in G$ satisfies

$$\dot U = -H'(Y(t))U, \quad U(0) = I. \tag{5.40}$$

Equation (5.39) can be written as

$$Y(t) = \mathrm{Ad}^*_{U(t)^{-1}} Y_0,$$

where $\mathrm{Ad}^*_{U^{-1}}$ is the adjoint of $\mathrm{Ad}_{U^{-1}}$. The solution $Y(t)$ of (5.37) thus lies on the *coadjoint orbit*

$$Y(t) \in \{\mathrm{Ad}^*_{U^{-1}} Y_0\,;\, U \in G\}. \tag{5.41}$$

In coordinates $Y = \sum_{j=1}^n y_j F_j$, we note $y_j = \langle Y_0, U(t)^{-1} E_j U(t)\rangle$.

Proof. Differentiating the ansatz (5.39) for the solution we obtain

$$\begin{aligned}\langle \dot{Y}, X\rangle &= \langle Y_0, -U^{-1}\dot{U}U^{-1}XU + U^{-1}X\dot{U}\rangle \\ &= \langle Y_0, U^{-1}[X, \dot{U}U^{-1}]U\rangle = \langle Y, [X, \dot{U}U^{-1}]\rangle,\end{aligned}$$

where we have used (5.39) in the first and the last equation. This shows that (5.37) is satisfied with the choice $\dot{U}U^{-1} = -H'(Y)$. □

Example 5.7 (Rigid Body). The Lie group corresponding to the rigid body is $\mathrm{SO}(3)$ with the Lie algebra $\mathfrak{so}(3)$ of skew-symmetric 3×3 matrices, with the basis

$$E_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

If we let $x = (x_1, x_2, x_3)^T$ be the coordinates of $X \in \mathfrak{so}(3)$, then we have $Xv = x \times v$ for all $v \in \mathbb{R}^3$. Since for $U \in \mathrm{SO}(3)$,

$$U^{-1}XUv = U^{-1}(x \times Uv) = U^{-1}x \times v,$$

the vector $U^{-1}x$ consists of the coordinates of $U^{-1}XU \in \mathfrak{so}(3)$.

Let $y = (y_1, y_2, y_3)^T$ be the coordinates of $Y \in \mathfrak{so}(3)^*$ with respect to the dual basis of E_1, E_2, E_3. Since

$$\langle Y, U^{-1}XU\rangle = \Big\langle \sum_{j=1}^{3} y_j F_j, \sum_{i=1}^{3} (U^{-1}x)_i E_i \Big\rangle = y^T U^{-1} x = (Uy)^T x,$$

the coordinates of $\mathrm{Ad}_{U^{-1}} Y$ are given by the vector Uy. Therefore, the coordinates of the coadjoint orbit of Y lie on a sphere of radius $\|y\|$. The conservation of the coadjoint orbit thus reduces here to the preservation of the Casimir $C(y) = y_1^2 + y_2^2 + y_3^2$.

Lie–Poisson integrators seem to be have first been considered by Ge & Marsden (1988), who extend the construction of symplectic methods by generating functions to Lie–Poisson systems. Channel & Scovel (1991) propose an implementation of these methods based on a coordinatization of the group by the exponential map.

McLachlan (1993) proposes integrators based on splitting the Hamiltonian and illustrates this approach for various examples of Lie-Poisson systems. When applicable, such splitting integrators yield Poisson integrators that preserve the coadjoint orbits, since they are composed of exact flows of Lie-Poisson systems.

Engø & Faltinsen (2001) propose to solve numerically the Lie–Poisson system (5.34) by applying Lie group integrators such as those of Sect. IV.8 to the differential equation (5.40) with (5.39). This approach keeps the solution on the coadjoint orbit by construction, but it does not, in general, give a Poisson integrator.

VII.5.5 Lie–Poisson Reduction

The reduction of the Hamiltonian equations of motion of the free rigid body to the Euler equations is an instance of a general reduction process from Hamiltonian systems with symmetry on a Lie group to Lie–Poisson systems, which we now describe; cf. Marsden & Ratiu (1999), Chap. 13, for a presentation in a more abstract framework and for an historical account.

Let us assume that the Lie group G is a subgroup of GL(n) given by

$$G = \{Q\,;\, g_i(Q) = 0,\; i = 1,\ldots,m\}, \tag{5.42}$$

and consider a Hamiltonian system on G,

$$\begin{aligned} \dot P &= -\nabla_Q H(P,Q) - \sum_{i=1}^{m} \lambda_i \nabla_Q g_i(Q), \qquad \dot Q = \nabla_P H(P,Q) \\ 0 &= g_i(Q), \quad i = 1,\ldots,m, \end{aligned} \tag{5.43}$$

where P,Q are square matrices, and $\nabla_Q H = \big(\partial H/\partial Q_{ij}\big)$. This is of the form discussed in Sect. VII.1.2. In regions where the matrix

$$\left(\frac{\partial^2 H(P,Q)}{\partial P^2}\Big(\nabla_Q g_i(Q), \nabla_Q g_j(Q)\Big)\right)_{i,j=1}^{m} \qquad \text{is invertible,} \tag{5.44}$$

the Lagrange parameters λ_i can be expressed in terms of P and Q (cf. formula (1.13)). Hence, a unique solution exists locally provided the initial values (P_0,Q_0) are consistent, i.e., $g_i(Q_0) = 0$ and

$$g_i'(Q_0)\Big(\nabla_P H(P_0,Q_0)\Big) = \text{trace}\,\Big(\nabla_Q g_i(Q_0)^T\, \nabla_P H(P_0,Q_0)\Big) = 0,$$

or equivalently, $Q_0 \in G$ and $\nabla_P H(P_0,Q_0) \in T_{Q_0}G$.

We now assume that the Hamiltonian H is quadratic in P. As we have seen in Sect. VII.1.2, the equations (5.43) can be viewed as a differential equation on the cotangent bundle $T^*G = \{(P,Q)\,;\, Q \in G,\; P \in T_Q^*G\}$, where the cotangent space T_Q^*G is identified with a subspace of matrices such that

$$P \in T_Q^*G \qquad \text{if and only if} \qquad \nabla_P H(P,Q) \in T_Q G. \tag{5.45}$$

With this identification, the duality beween T_Q^*G and T_QG is given by the matrix inner product

$$\langle P, V\rangle = \text{trace}\,(P^T V) \qquad \text{for } P \in T_Q^*G,\; V \in T_QG.$$

We call the Hamiltonian *left-invariant*, if

$$H(U^T P, U^{-1}Q) = H(P,Q) \qquad \text{for all } U \in G. \tag{5.46}$$

In this case we have $H(P,Q) = H(Q^T P, I)$ and by differentiating we obtain $\nabla_P H(P,Q) = Q\nabla_P H(Q^T P, I)$. By (5.45) and since $T_Q G = \{QX\,;\, X \in \mathfrak{g}\}$ with the Lie algebra $\mathfrak{g} = T_I G$ (cf. Sect. IV.6), this relation implies

$$P \in T_Q^* G \qquad \text{if and only if} \qquad Q^T P \in T_I^* G = \mathfrak{g}^*. \tag{5.47}$$

Now $H(P,Q)$ depends, for $(P,Q) \in T^*G$, only on the product $Y = Q^T P \in \mathfrak{g}^*$, and we write[7] $H(P,Q) = H(Y)$ with a function $H : \mathfrak{g}^* \to \mathbb{R}$.

Left-invariant Hamiltonian systems can be reduced to a Lie–Poisson system of half the dimension by a process that generalizes the derivation of the Euler equations for the rigid body.

Theorem 5.8. *Consider a Hamiltonian system (5.43) on a matrix Lie group G with a left-invariant quadratic Hamiltonian $H(P,Q) = H(Y)$ for $Y = Q^T P$. If $\big(P(t), Q(t)\big) \in T^*G$ is a solution of the system (5.43), then $Y(t) = Q(t)^T P(t) \in \mathfrak{g}^*$ solves the differential equation (5.37).*

Proof. It is convenient for the proof (though not necessary, see the lines following (2.17)) to extend the Hamiltonian $H : \mathfrak{g}^* \to \mathbb{R}$ to a function of arbitrary matrices Y by setting $H(Y) = H(\Pi Y)$, where Π is the projection onto $\mathfrak{g}^*$ given by $\langle \Pi Y, X\rangle = (Y, X)$ for all $X \in \mathfrak{g}$, with the matrix inner product $(Y,X) = \text{trace}\,(Y^T X)$.

We first compute the derivatives of $H(P,Q) = H(Y) = H(\Pi Y) = H(y)$ where $Q^T P = Y$ and, using the notation of Theorem 5.5, $\Pi Y = \sum_{j=1}^d y_j F_j$. Since $y_j = \langle \Pi Q^T P, E_j\rangle = (Q^T P, E_j)$, it follows from $\nabla_A \,\text{trace}\,(A^T B) = B$ that

$$\nabla_P H(P,Q) = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j}\, \nabla_P y_j = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j}\, \nabla_P \,\text{trace}\,(P^T Q E_j) = Q H'(Y), \tag{5.48}$$

where $H'(Y) = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j} E_j \in \mathfrak{g}$ as in Theorem 5.5. Using the identity $y_j = \text{trace}\,(P^T Q E_j) = \text{trace}\,(Q^T P E_j^T)$ we get in a similar way

$$\nabla_Q H(P,Q) = P H'(Y)^T. \tag{5.49}$$

Consequently, the differential equations (5.43) become

$$\dot P \;=\; -PH'(Q^T P)^T - \sum_{i=1}^m \lambda_i \,\nabla_Q g_i(Q), \qquad \dot Q \;=\; QH'(Q^T P)\,. \tag{5.50}$$

The product rule $\dot Y = \dot Q^T P + Q^T \dot P$ for $Y = Q^T P$ thus yields

$$\dot Y \;=\; H'(Y)^T Y - Y\, H'(Y)^T - \sum_{i=1}^m \lambda_i \, Q^T \nabla_Q g_i(Q). \tag{5.51}$$

[7] We use again the same letter for different functions. Since they have either one or two arguments, no confusion should arise.

For $X \in \mathfrak{g}$, we now exploit the properties

$$\begin{aligned} \big\langle Q^T \nabla_Q g_i(Q), X \big\rangle &= \big\langle \nabla_Q g_i(Q), QX \big\rangle = 0 \qquad \text{(because } QX \in T_Q G) \\ \big\langle [H'(Y)^T, Y], X \big\rangle &= \text{trace}\, \big(\big(Y^T H'(Y) - H'(Y) Y^T \big) X \big) \\ &= \text{trace}\, \big(Y^T \big(H'(Y) X - X H'(Y) \big) \big) = \big\langle Y, [H'(Y), X] \big\rangle. \end{aligned}$$

Since $Y(t) \in \mathfrak{g}^*$ for all t, this gives the equation (5.37). □

Reduced System and Reconstruction. Combining Theorems 5.8 and 5.5, we have reduced the Hamiltonian system (5.43) to the Lie-Poisson system for $y(t) \in \mathbb{R}^d$,

$$\dot{y} = B(y) \nabla H(y), \tag{5.52}$$

of half the dimension. To recover the full solution $(P(t), Q(t)) \in T^*G$, we must solve this system along with

$$\dot{Q} = Q H'(Y)\,, \qquad P = Q^{-T} Y \tag{5.53}$$

where $Y = \sum_{j=1}^d y_j F_j \in \mathfrak{g}^*$.

Poisson Structures. The Poisson bracket on $\mathbb{R}^d$ defined by $B(y)$ is still closely related to the canonical Poisson bracket on $\mathbb{R}^{2n^2}$. Consider left-invariant real-valued functions K, L on T^*G. These can be viewed as functions on $T^*G/G = \mathfrak{g}^* \subset \mathbb{R}^{n\times n}$,

$$K(P,Q) = K(Y) \qquad \text{for} \quad Y = Q^T P.$$

(As we did previously in this section, we use the same symbol for these functions.) Via the projection $\Pi : \mathbb{R}^{n\times n} \to \mathfrak{g}^*$ used in the above proof, we can extend $K(Y) = K(\Pi Y)$ to arbitrary $n \times n$ matrices Y, and via the above relation to a left-invariant function $K(P,Q)$ on $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$, on which we have the canonical Poisson bracket

$$\{K, L\}_{\text{can}} = \sum_{k,l=1}^{n} \Big(\frac{\partial K}{\partial Q_{kl}} \frac{\partial L}{\partial P_{kl}} - \frac{\partial K}{\partial P_{kl}} \frac{\partial L}{\partial Q_{kl}} \Big).$$

On the other hand, we can view K as a function on $\mathbb{R}^d$ by choosing coordinates on $\mathfrak{g}^*$,

$$K(y) = K(Y) \qquad \text{for} \quad Y = \sum_{j=1}^{d} y_j F_j \in \mathfrak{g}^*.$$

On $\mathbb{R}^d$ we have the Poisson bracket defined by the structure matrix $B(y)$,

$$\{K, L\} = \sum_{i,j=1}^{d} \frac{\partial K}{\partial y_i}\, b_{ij}\, \frac{\partial L}{\partial y_j}\,.$$

Lemma 5.9. *For left-invariant functions K, L as described above, we have for $Q^T P = Y = \sum_{j=1}^d y_j F_j \in \mathfrak{g}^*$*

$$\{K, L\}(y) = \big\langle Y, [L'(Y), K'(Y)] \big\rangle = \{K, L\}_{\mathrm{can}}(P, Q)$$

where $K'(Y) = \sum_{i=1}^d \frac{\partial K}{\partial y_i}(y) E_i \in \mathfrak{g}$.

Proof. The first equality follows from the identity

$$b_{ij}(y) = \big\langle Y, [E_j, E_i] \big\rangle ,$$

which is a direct consequence of the definition (5.35) with (5.36). For the second equality, the relations (5.48) and (5.49) for K and L yield

$$\begin{aligned} \{K, L\}_{\mathrm{can}}(P, Q) &= \mathrm{trace}\,\big(K'(Y)Y^T L'(Y) - K'(Y)^T Y L'(Y)^T\big) \\ &= \mathrm{trace}\,\big(K'(Y)Y^T L'(Y) - L'(Y)Y^T K'(Y)\big) \\ &= \mathrm{trace}\,\big(Y^T [L'(Y), K'(Y)]\big) = \big\langle Y, [L'(Y), K'(Y)] \big\rangle , \end{aligned}$$

which is the result. □

Discrete Lie–Poisson Reduction. Consider a symplectic integrator

$$(P_1, Q_1) = \Phi_h(P_0, Q_0) \qquad \text{on } \ T^*G$$

for the left-invariant Hamiltonian system (5.43), and suppose that the method preserves the left-invariance: if $\Phi_h(P_0, Q_0) = (P_1, Q_1)$, then

$$\Phi_h(U^T P_0, U^{-1} Q_0) = (U^T P_1, U^{-1} Q_1) \quad \text{for all } \ U \in G. \tag{5.54}$$

For example, this is satisfied by the RATTLE algorithm. The method then induces a one-step map

$$Y_1 = \Psi_h(Y_0) \qquad \text{on } \ \mathfrak{g}^*$$

by setting $Y_1 = Q_1^T P_1$ for $(P_1, Q_1) = \Phi_h(P_0, Q_0)$ with $Q_0^T P_0 = Y_0$. This is a numerical integrator for (5.37), and in the coordinates $y = (y_j)$ with respect to the basis (F_j) of $\mathfrak{g}^*$ this gives a map

$$y_1 = \psi_h(y_0) \qquad \text{on } \ \mathbb{R}^d,$$

which is a numerical integrator for the Poisson system (5.52).

Example 5.10. For the rigid body, applying the RATTLE algorithm to the constrained Hamiltonian system (5.18) yields the integrator for the Euler equations discussed in Sect. VII.5.3. By the following result this is a Poisson integrator.

Theorem 5.11. *If $\Phi_h(P, Q)$ is a symplectic and left-invariant integrator for (5.43), then its reduction $\psi_h(y)$ is a Poisson map.*

Proof. We write ψ_h as the composition

$$\psi_h : \mathbb{R}^d \xrightarrow{\xi} T^*G \xrightarrow{\Phi_h} T^*G \xrightarrow{\eta} \mathbb{R}^d$$

where $\eta = (\eta_j)$ is the function with $\eta_j(P,Q) = y_j$ for $Q^T P = \sum_{j=1}^d y_j F_j$, and ξ is any right inverse of η, i.e., $\eta \circ \xi = \mathrm{id}$. For arbitrary smooth real-valued functions K, L on $\mathbb{R}^d$ we then have for $(P,Q) = \xi(y)$, using Lemma 5.9 in the outer equalities and the symplecticity of Φ_h in the middle equality,

$$\begin{aligned} \{K \circ \psi_h, L \circ \psi_h\}(y) &= \{K \circ \eta \circ \Phi_h, L \circ \eta \circ \Phi_h\}_{\mathrm{can}}(P,Q) \\ &= \{K \circ \eta, L \circ \eta\}_{\mathrm{can}}\big(\Phi_h(P,Q)\big) = \{K, L\}\big(\psi_h(y)\big). \end{aligned}$$

This equation states that ψ_h is a Poisson map. □

A similar reduction in a discrete Lagrangian framework is studied by Marsden, Pekarsky & Shkoller (1999).

The reduced numerical maps ψ_h and Ψ_h have further structure-preserving properties: they preserve the Casimirs and the co-adjoint orbits. This will be shown in Sect. IX.5.3 with the help of backward error analysis.

VII.6 Reduced Models of Quantum Dynamics

To incorporate quantum effects in molecular dynamics simulations, computations are done with models that are intermediate between classical molecular dynamics based on Newton's equations of motion and full quantum dynamics described by the N-particle Schrödinger equation. The direct computational treatment of the latter is not feasible because of its high dimensionality. These intermediate models are obtained by the Hamiltonian reduction (2.17) from an infinite-dimensional Hilbert space to an appropriately chosen manifold. In chemical physics, this reduction is known as the *Dirac–Frenkel time-dependent variational principle*. We illustrate this procedure for the case where the quantum-mechanical wave function is approximated by a complex Gaussian as proposed by Heller (1975). It turns out that the resulting ordinary differential equations have a Poisson structure, which was recently described by Faou & Lubich (2004). Following that paper, we derive a structure-preserving explicit integrator for Gaussian wavepackets, which tends to the Störmer–Verlet method in the classical limit.

VII.6.1 Hamiltonian Structure of the Schrödinger Equation

> The introduction of wave mechanics stands ... as Schrödinger's monument and a worth one. (From Schrödinger's obituary in *The Times* 1961; quoted from http://www-groups.dcs.st-and.ac.uk/ history/Mathematicians/Schrodinger.html)

The time-dependent N-body Schrödinger equation reads (see, e.g., Messiah (1999) and Thaller (2000))

$$i\varepsilon \frac{\partial \psi}{\partial t} = H\psi \tag{6.1}$$

for the wave function $\psi = \psi(x,t)$ depending on the spatial variables $x = (x_1, \dots, x_N)$ with $x_k \in \mathbb{R}^d$ (e.g., with $d = 1$ or 3 in the partition) and the time $t \in \mathbb{R}$. The squared absolute value $|\psi(x,t)|^2$ represents the joint probability density for N particles to be at the positions $x_1, \dots, x_N$ at time t. In (6.1), ε is a (small) positive number representing the scaled Planck constant and i is the complex imaginary unit. The Hamiltonian operator H is written

$$H = T + V$$

with the kinetic and potential energy operators

$$T = -\sum_{k=1}^{N} \frac{\varepsilon^2}{2m_k} \Delta_{x_k} \quad \text{and} \quad V = V(x),$$

where $m_k > 0$ is a particle mass and Δ_{x_k} the Laplacian in the variable $x_k \in \mathbb{R}^d$, and where the real-valued potential V acts as a multiplication operator $(V\phi)(x) = V(x)\phi(x)$. Under appropriate conditions on V (boundedness of V is sufficient, but by no means necessary), the operator H is then a self-adjoint operator on the complex Hilbert space $L^2(\mathbb{R}^{dN}, \mathbb{C})$ with domain $D(H) = D(T) = \{\phi \in L^2(\mathbb{R}^{dN}, \mathbb{C})\,;\, T\phi \in L^2(\mathbb{R}^{dN}, \mathbb{C})\}$; see Sect. V.5.3 of Kato (1980).

We separate the real and imaginary parts of $\psi = v + iw \in L^2(\mathbb{R}^{dN}, \mathbb{C})$, the complex Hilbert space of Lebesgue square-integrable functions. The functions v and w are thus functions in the *real* Hilbert space $L^2(\mathbb{R}^{dN}, \mathbb{R})$. We denote the complex inner product by $\langle \cdot, \cdot \rangle$ and the real inner product by $(\cdot, \cdot)$. The L^2 norms will be simply denoted by $\| \cdot \|$.

As H is a real operator, formula (6.1) can be written

$$\begin{aligned} \varepsilon \dot{v} &= Hw, \\ \varepsilon \dot{w} &= -Hv, \end{aligned} \tag{6.2}$$

or equivalently, with the canonical structure matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and the Hamiltonian function (we use the same symbol H as for the operator)

$$H(v,w) = \frac{1}{2}\langle \psi, H\psi \rangle = \frac{1}{2}(v, Hv) + \frac{1}{2}(w, Hw)$$

for $\psi = v + iw$ in the domain of the operator H. This becomes the canonical Hamiltonian system

$$\begin{pmatrix} \dot{v} \\ \dot{w} \end{pmatrix} = \varepsilon^{-1} J^{-1} \nabla H(v,w).$$

Note that the real multiplication with J corresponds to the complex multiplication with the imaginary unit i. The flow of this system preserves the canonical symplectic two-form

$$\omega(\xi_1,\xi_2) = (J\xi_1,\xi_2), \qquad \xi_1,\xi_2 \in L^2(\mathbb{R}^{dN},\mathbb{R})^2. \tag{6.3}$$

VII.6.2 The Dirac–Frenkel Variational Principle

> For dealing with atoms involving many electrons the accurate quantum theory, involving a solution of the wave equation in many-dimensional space, is far too complicated to be practicable. One must therefore resort to approximate methods. (P.A.M. Dirac 1930)

Reduced models of the Schrödinger equation are obtained by restricting the equation to an approximation manifold $\mathcal{M}$ via (2.17), viz.,

$$(\varepsilon J\dot{u} - \nabla H(u),\, \xi) = 0 \qquad \text{for all} \quad \xi \in T_u\mathcal{M}, \tag{6.4}$$

or equivalently in complex notation for $u = (v,w)^T = v + iw$,

$$\mathrm{Re}\,\langle\, \varepsilon i\dot{u} - Hu,\, \xi\,\rangle = 0 \qquad \text{for all} \quad \xi \in T_u\mathcal{M}. \tag{6.5}$$

Taking the real part can be omitted if the tangent space $T_u\mathcal{M}$ is complex linear. Equation (6.5) (usually without the real part) is known as the Dirac–Frenkel time-dependent variational principle, after Dirac (1930) and Frenkel (1934); see also McLachlan (1964), Heller (1976), Beck, Jäckle, Worth & Meyer (2000), and references therein.

We choose a (local) coordinate map $u = \chi(y)$ of $\mathcal{M}$ and denote its derivative $X_{\mathbb{C}}(y) = V(y) + iW(y) = \chi'(y)$ or in the real setting as $X = \begin{pmatrix} V \\ W \end{pmatrix}$. Denoting by X^T the adjoint of X with respect to the real inner product $(\cdot,\cdot)$, we thus obtain

$$\varepsilon X(y)^T J X(y)\, \dot{y} = X(y)^T \nabla_u H(\chi(y)).$$

With $X_{\mathbb{C}}^*$ denoting the adjoint of $X_{\mathbb{C}}$ with respect to the complex inner product $\langle\cdot,\cdot\rangle$, we note $X_{\mathbb{C}}^* X_{\mathbb{C}} = (V^TV + W^TW) + i(V^TW - W^TV) = X^TX - iX^TJX$ and hence

$$X^TJX = -\mathrm{Im}\, X_{\mathbb{C}}^* X_{\mathbb{C}}. \tag{6.6}$$

Lemma 6.1. *If $T_u\mathcal{M}$ is a complex linear space for every $u \in \mathcal{M}$, then $\mathcal{M}$ is a symplectic submanifold of $L^2(\mathbb{R}^N,\mathbb{R})^2$, that is, the symplectic two-form (6.3) is non-degenerate on $T_u\mathcal{M}$ for all $u \in \mathcal{M}$. Expressed in coordinates,*

$$X(y)^T J X(y) \ \ \textit{is invertible for all } y.$$

Proof. We fix $u = \chi(y) \in \mathcal{M}$ and omit the argument y in the following. Since $T_u\mathcal{M} = \mathrm{Range}(X_{\mathbb{C}})$ is complex linear by assumption, there exists a real linear mapping $L : \mathbb{R}^m \to \mathbb{R}^m$ such that $iX_{\mathbb{C}}\eta = X_{\mathbb{C}}L\eta$ for all $\eta \in \mathbb{R}^m$. This implies

$$JX = XL \quad \text{and} \quad L^2 = -\mathrm{Id}$$

and hence $X^TJX = X^TXL$, which is invertible. □

Approximation properties of the Dirac–Frenkel variational principle can be obtained from the interpretation as the *orthogonal* projection $\dot{u} = P_\perp(u)\frac{1}{i\varepsilon}Hu$, which corresponds to taking the imaginary part in (6.5), as opposed to the symplectic projection in (6.4) which corresponds to the real part. See Lubich (2005) for a near-optimality result for approximation on the manifold.

VII.6.3 Gaussian Wavepacket Dynamics

> We develop a new approach to semiclassical dynamics which exploits the fact that extended wavefunctions for heavy particles (or particles in harmonic potentials) may be decomposed into time-dependent wave packets, which spread minimally and which execute classical or nearly classical trajectories. A Gaussian form for the wave packets is assumed and equations of motion are derived for the parameters characterizing the Gaussian. (E.J. Heller 1975)

The variational Gaussian wavepacket dynamics of Heller (1976) is obtained by choosing the manifold $\mathcal{M}$ in (6.5) as consisting of complex Gaussians. For ease of presentation we restrict our attention in the following to the one-particle case $N = 1$ (the extension to $N > 1$ is straightforward; cf. Heller (1976) and Faou & Lubich (2004)). Here we have

$$\mathcal{M} = \{u = \chi(y) \in L^2(\mathbb{R}^d, \mathbb{C}) : y = (p, q, \alpha, \beta, \gamma, \delta) \in \mathbb{R}^{2d+4} \text{ with } \beta > 0\} \tag{6.7}$$

with

$$\Big(\chi(y)\Big)(x) = \exp\Big(\frac{i}{\varepsilon}\big((\alpha + i\beta)\,|x - q|^2 + p \cdot (x - q) + \gamma + i\delta\big)\Big), \tag{6.8}$$

where $|\cdot|$ and $\cdot$ stand for the Euclidean norm and inner product on $\mathbb{R}^d$. The parameters q and p represent the average position and momentum, respectively: for $u = \chi(y)$ with $y = (p, q, \alpha, \beta, \gamma, \delta)$ and $\|u\| = 1$, a direct calculation shows that

$$q = \langle\, u, xu \,\rangle = \int_{\mathbb{R}^d} x\,|u(x)|^2\,dx\,, \qquad p = \langle\, u, -i\varepsilon\nabla u \,\rangle.$$

The parameter $\beta > 0$ determines the width of the wavepacket. The tangent space $T_u\mathcal{M} \subset L^2(\mathbb{R}^d, \mathbb{C})$ at a given point $u = \chi(y) \in \mathcal{M}$ is $(2d + 4)$-dimensional and is made of the elements of $L^2(\mathbb{R}^d, \mathbb{C})$ written as

$$\frac{i}{\varepsilon}\Big((A + iB)\,|x - q|^2 + (P - 2(\alpha + i\beta)Q) \cdot (x - q) - p \cdot Q + C + iD\Big)u \tag{6.9}$$

with arbitrary $(P, Q, A, B, C, D)^T \in \mathbb{R}^{2d+4}$. We note that $T_u\mathcal{M}$ is complex linear, and $u \in T_u\mathcal{M}$. By choosing $\xi = iu$ in (6.5), this yields $(d/dt)\|u\|^2 = 2\,\mathrm{Re}\,\langle\dot{u}, u\rangle = 0$ and hence the preservation of the squared L^2 norm of $u = \chi(y)$, which is given by

$$
\begin{aligned}
I(y) &= \|u\|^2 = \int_{\mathbb{R}^d} |u(x)|^2 \, dx \qquad (6.10) \\
&= \int_{\mathbb{R}^d} \exp\Bigl(-\frac{2}{\varepsilon}\bigl(\beta|x-q|^2+\delta\bigr)\Bigr)\, dx = \exp\left(-\frac{2\delta}{\varepsilon}\right)\left(\frac{\pi\varepsilon}{2\beta}\right)^{d/2}.
\end{aligned}
$$

The physically reasonable situation is $\|u\|^2 = 1$, which corresponds to the interpretation of $|u(x)|^2$ as a probability density.

With these preparations we have the following result of Faou & Lubich (2004).

Theorem 6.2. *The Hamiltonian reduction of the Schrödinger equation to the Gaussian wavepacket manifold $\mathcal{M}$ of (6.7)-(6.8) yields the Poisson system*

$$
\dot{y} = B(y)\nabla K(y) \qquad (6.11)
$$

where, for $y = (p, q, \alpha, \beta, \gamma, \delta) \in \mathbb{R}^{2d+4}$ with $\beta > 0$, and with 1_d denoting the d-dimensional identity,

$$
B(y) = \frac{1}{I(y)} \begin{pmatrix}
0 & -1_d & 0 & 0 & -p & 0 \\
1_d & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{4\beta^2}{\varepsilon d} & 0 & -\beta \\
0 & 0 & -\frac{4\beta^2}{\varepsilon d} & 0 & \beta & 0 \\
p^T & 0 & 0 & -\beta & 0 & \frac{d+2}{4}\varepsilon \\
0 & 0 & \beta & 0 & -\frac{d+2}{4}\varepsilon & 0
\end{pmatrix} \qquad (6.12)
$$

defines a Poisson structure, and for $u = \chi(y)$,

$$
K(y) = \langle u, Hu \rangle = K_T(y) + K_V(y) \qquad (6.13)
$$

is the total energy, with kinetic and potential parts

$$
K_T(y) = I(y)\left(\frac{|p|^2}{2m} + \frac{\varepsilon d}{2m}\,\frac{\alpha^2+\beta^2}{\beta}\right) = \langle u, Tu \rangle
$$

and

$$
K_V(y) = \int_{\mathbb{R}^d} V(x) \exp\Bigl(-\frac{2}{\varepsilon}\bigl(\beta|x-q|^2+\delta\bigr)\Bigr)\, dx = \langle u, Vu \rangle.
$$

Both $K(y)$ and $I(y)$ are first integrals of the system.

Proof. As in (2.22), the differential equation for y is $\varepsilon X(y)^T J X(y)\dot{y} = \frac{1}{2}\nabla K(y)$. We note (6.6) and

$$
X_{\mathbb{C}}(y) = \frac{i}{\varepsilon}\bigl(x-q\,,\; -2a(x-q)-p\,,\; |x-q|^2\,,\; i|x-q|^2\,,\; 1\,,\; i\bigr)\,u
$$

where $a = \alpha + i\beta$ and $u = \chi(y)$ in the complex setting. Using the calculus of Gaussian integrals, we compute

$$\varepsilon X^T(y)JX(y) = \frac{1}{2}\,I(y)\begin{pmatrix} 0 & 1_d & 0 & 0 & 0 & 0 \\ -1_d & 0 & 0 & \frac{dp}{2\beta} & 0 & \frac{2p}{\varepsilon} \\ 0 & 0 & 0 & -\frac{\varepsilon d(d+2)}{8\beta^2} & 0 & -\frac{d}{2\beta} \\ 0 & -\frac{dp^T}{2\beta} & \frac{\varepsilon d(d+2)}{8\beta^2} & 0 & \frac{d}{2\beta} & 0 \\ 0 & 0 & 0 & -\frac{d}{2\beta} & 0 & -\frac{2}{\varepsilon} \\ 0 & -\frac{2p^T}{\varepsilon} & \frac{d}{2\beta} & 0 & \frac{2}{\varepsilon} & 0 \end{pmatrix},$$

and inversion yields the differential equation with $B(y) = (2\varepsilon X^T(y)JX(y))^{-1}$ as stated. The system is a Poisson system by Theorem 2.8. □

Assuming $I(y) = \|u\|^2 = 1$, we observe that the differential equations for the average position and momentum, q and p, read

$$\dot q = p/m\,, \qquad \dot p = -\langle\, u, \nabla V u\,\rangle \tag{6.14}$$

for $u = \chi(y)$ and $y = (p, q, \alpha, \beta, \gamma, \delta)$. We then note $\langle\, u, \nabla V u\,\rangle \to \nabla V(q)$ as $\varepsilon \to 0$. The differential equations for q and p thus tend to Newtonian equations of motion in the classical limit $\varepsilon \to 0$:

$$\dot q = p/m\,, \qquad \dot p = -\nabla V(q). \tag{6.15}$$

It will be useful to consider also scaled variables

$$\widehat y = (p, q, \alpha, \widehat\beta, \gamma, \widehat\delta) \in \mathbb{R}^{2d+4} \qquad \text{with} \quad \widehat\beta = \frac{\beta}{\varepsilon},\ \widehat\delta = \frac{\delta}{\varepsilon}. \tag{6.16}$$

Here we have

$$\dot{\widehat y} = \widehat B(\widehat y)\nabla \widehat K(\widehat y) \tag{6.17}$$

where the structure matrix $\widehat B(\widehat y)$ is independent of ε, and where $\widehat K(\widehat y)$ depends regularly on $\varepsilon \ge 0$.

VII.6.4 A Splitting Integrator for Gaussian Wavepackets

With the natural splitting $H = T + V$ into kinetic and potential energy, we now consider the variational splitting integrator (4.7) – (4.8), which here becomes the following.

1. We define u_n^+ in $\mathcal{M}$ as the solution at time $h/2$ of the equation for u,

$$\langle\, i\varepsilon \dot u - Vu, \xi\,\rangle = 0 \quad \text{for all } \ \xi \in T_u\mathcal{M} \tag{6.18}$$

   with initial value $u(0) = u_n \in \mathcal{M}$.
2. We define u_{n+1}^- as the solution at time h of

$$\langle\, i\varepsilon \dot u - Tu, \xi\,\rangle = 0 \quad \text{for all } \ \xi \in T_u\mathcal{M} \tag{6.19}$$

   with initial value $u(0) = u_n^+$.
3. Then u_{n+1} is the solution at time $h/2$ of (6.18) with initial value $u(0) = u_{n+1}^-$.

By Theorem 6.2, the substeps in the definition of this splitting method written in the coordinates $y = (p, q, \alpha, \beta, \gamma, \delta)$ are the exact flows $\varphi_{h/2}^V$ and φ_h^T of the Poisson systems

$$\dot{y} = B(y)\nabla K_V(y) \quad \text{and} \quad \dot{y} = B(y)\nabla K_T(y).$$

Note that both equations preserve the L^2 norm of $u = \chi(y)$, which we assume to be 1 in the following.

Most remarkably, these equations can be solved explicitly. Let us consider first the equations (6.19). They are written, for $a = \alpha + i\beta$ and $c = \gamma + i\delta$, as

$$\left\{ \begin{array}{rcl} \dot{q} &=& p/m, \\ \dot{p} &=& 0, \end{array} \right. \qquad \left\{ \begin{array}{rcl} \dot{a} &=& -2a^2/m, \\ \dot{c} &=& \big(\frac{1}{2}|p|^2 + i\varepsilon d a\big)/m, \end{array} \right. \tag{6.20}$$

with initial values $y_0 = (p_0, q_0, a_0, c_0)$ corresponding to $u_0 = \chi(y_0)$. They have the solution

$$q(t) = q_0 + \frac{t}{m}\,p_0, \qquad p(t) = p_0, \qquad a(t) = \frac{a_0}{1 + 2a_0 t/m},$$

and

$$c(t) = c_0 + t\,\frac{|p_0|^2}{2m} + \frac{i\varepsilon d}{2}\log\left(1 + \frac{2a_0 t}{m}\right).$$

Let us now consider the equations (6.18). Taking into account the fact that the potential V is real, these equations are written

$$\begin{array}{ll} \dot{p} = -\langle\, u, \nabla V u\,\rangle, & \dot{q} = 0, \\ \dot{\alpha} = -\frac{1}{2d}\langle\, u, \varDelta V u\,\rangle, & \dot{\beta} = 0, \\ \dot{\gamma} = -\langle\, u, V u\,\rangle + \frac{\varepsilon}{8\beta}\langle\, u, \varDelta V u\,\rangle, & \dot{\delta} = 0, \end{array} \tag{6.21}$$

with the L^2 inner products

$$\langle\, u, W u\,\rangle = \int_{\mathbb{R}^d} W(x)\,\exp\Big(-\frac{2}{\varepsilon}\big(\beta|x-q|^2 + \delta\big)\Big)\,dx \tag{6.22}$$

for $W = V, \nabla V, \varDelta V$. As the L^2 inner products in the equations for p, α, γ depend only on q, β, δ which are constant along this trajectory, these equations can be solved trivially, requiring only the computation of the inner products at the initial value. We thus see that the splitting scheme $\varPhi_h = \varphi_{h/2}^V \circ \varphi_h^T \circ \varphi_{h/2}^V$ can be computed explicitly. This gives the following algorithm (Faou & Lubich 2004).

Algorithm 6.3 (Gaussian Wavepacket Integrator). *A step from time t_n to t_{n+1}, starting from the Gaussian wavepacket $u_n = \chi(p_n, q_n, \alpha_n, \beta_n, \gamma_n, \delta_n)$, proceeds as follows:*

1. *With $\langle W\rangle_n = \langle\, u_n, W u_n\,\rangle$ given by (6.22) for $W = V, \nabla V, \varDelta V$, compute*

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\,\langle \nabla V \rangle_n \\ \alpha_n^+ &= \alpha_n - \frac{h}{4d}\,\langle \Delta V \rangle_n \\ \gamma_n^+ &= \gamma_n + \frac{h\,\varepsilon}{16\beta_n}\,\langle \Delta V \rangle_n. \end{aligned} \tag{6.23}$$

2. *From the values* $p_{n+1/2}$, $a_n^+ = \alpha_n^+ + i\beta_n$ *and* $c_n^+ = \gamma_n^+ + i\delta_n$ *compute* q_{n+1}, $a_{n+1}^- = \alpha_{n+1}^- + i\beta_{n+1}$, *and* $c_{n+1}^- = \gamma_{n+1}^- + i\delta_{n+1}$ *via*

$$\begin{aligned} q_{n+1} &= q_n + \frac{h}{m}\,p_{n+1/2} \\ a_{n+1}^- &= a_n^+ \Big/ \Big(1 + \frac{2h}{m}\,a_n^+\Big) \\ c_{n+1}^- &= c_n^+ + \frac{i\varepsilon d}{2}\log\Big(1 + \frac{2h}{m}\,a_n^+\Big). \end{aligned} \tag{6.24}$$

3. *Compute* $p_{n+1}, \alpha_{n+1}, \gamma_{n+1}$ *from*

$$\begin{aligned} p_{n+1} &= p_{n+1/2} - \frac{h}{2}\,\langle \nabla V \rangle_{n+1} \\ \alpha_{n+1} &= \alpha_{n+1}^- - \frac{h}{4d}\,\langle \Delta V \rangle_{n+1} \\ \gamma_{n+1} &= \gamma_{n+1}^- + \frac{h\,\varepsilon}{16\beta_{n+1}}\,\langle \Delta V \rangle_{n+1}. \end{aligned} \tag{6.25}$$

Let us collect properties of this algorithm.

Theorem 6.4. *The splitting scheme of Algorithm 6.3 is an explicit, symmetric, second-order numerical method for Gaussian wavepacket dynamics (6.11)–(6.13). It is a Poisson integrator for the structure matrix (6.12), and it preserves the unit* L^2 *norm of the wavepackets:* $\|u_n\| = 1$ *for all* n.

In the limit $\varepsilon \to 0$, *the position and momentum approximations* q_n, p_n *of this method tend to those obtained by applying the Störmer–Verlet method to the associated classical mechanical system (6.15).*

The statement for $\varepsilon \to 0$ follows directly from the equations for $p_{n+1/2}$, q_{n+1}, p_{n+1} and from noting $\langle \nabla V \rangle_n \to \nabla V(q_n)$.

In view of the small parameter ε, the discussion of the order of the method requires more care. Here it is useful to consider the integrator in the scaled variables $\widehat{y} = (p, q, \alpha, \beta/\varepsilon, \gamma, \delta/\varepsilon)$ of (6.16). Since the differential equation (6.17) contains ε only as a regular perturbation parameter, after n steps of the splitting integrator we have the ε-uniform error bound

$$\widehat{y}_n - \widehat{y}(t_n) = O(h^2),$$

where the constants symbolized by the O-notation are independent of ε and of n and h with $nh \le \mathrm{Const}$. For the approximation of the absolute values of the Gaussian wavepackets this yields

$$\big\| |u_n|^2 - |u(t_n)|^2 \big\| = O(h^2), \tag{6.26}$$

but the approximation of the phases is only such that

$$\| u_n - u(t_n) \| = O(h^2/\varepsilon). \tag{6.27}$$

We refer to Faou & Lubich (2004) for the formulation of the corresponding algorithm for $N > 1$ particles, for further properties such as the exact conservation of linear and angular momentum and the long-time near-conservation of the total energy $\langle u_n, H u_n \rangle$, and for numerical experiments.

VII.7 Exercises

1. Prove that the Poisson bracket (2.8) satisfies the Jacobi identity (2.4) for all functions F, G, H, if and only if it satisfies (2.4) for the coordinate functions y_i, y_j, y_k.
   Hint (F. Engel, in Lie's *Gesammelte Abh.* vol. 5, p. 753). If the Jacobi identity is written as in (3.3), we see that there are no second partial derivatives of F (the left hand side is a Lie bracket, the right-hand side has no second derivatives of F anyway). Other permutations show the same result for G and H.
2. For x in an open subset of $\mathbb{R}^m$, let $A(x) = \big(a_{ij}(x)\big)$ be an invertible skew-symmetric $m \times m$-matrix, with

$$\frac{\partial a_{ij}}{\partial x_k} + \frac{\partial a_{ki}}{\partial x_j} + \frac{\partial a_{jk}}{\partial x_i} = 0 \qquad \text{for all } \ i, j, k. \tag{7.1}$$

   (a) Show that $B(x) = A(x)^{-1}$ satisfies (2.10) and hence defines a Poisson bracket.
   (b) Generalize Theorem 2.8 to Hamiltonian equations (2.18) with the two-form $\omega_x(\xi_1, \xi_2) = \xi_1^T A(x) \xi_2$.
   Remark. Condition (7.1) says that ω is a closed differential form.
3. Solve the following first order partial differential equation:

$$3\,\frac{\partial F}{\partial y_1} + 2\,\frac{\partial F}{\partial y_2} - 5\,\frac{\partial F}{\partial y_3} = 0.$$

   Result. $\quad f(2y_1 - 3y_2, 5y_2 + 2y_3)$.
4. Find two solutions of the homogeneous system

$$3\,\frac{\partial F}{\partial y_1} + \frac{\partial F}{\partial y_2} - 2\,\frac{\partial F}{\partial y_3} - 5\,\frac{\partial F}{\partial y_4} = 0, \qquad 2\,\frac{\partial F}{\partial y_1} - \frac{\partial F}{\partial y_2} - 3\,\frac{\partial F}{\partial y_4} = 0,$$

   such that their gradients are linearly independent.
5. Consider a Poisson system $\dot y = B(y)\nabla H(y)$ and a change of coordinates $z = \vartheta(y)$. Prove that in the new coordinates the system is of the form $\dot z = \widetilde B(z)\nabla K(z)$, where $\widetilde B(z) = \vartheta'(y)B(y)\vartheta'(y)^T$ (cf. formula (3.12)) and $K(z) = H(y)$.

6. Give an elementary proof of Theorem 4.3.
   Hint. Define $\delta(t) := \varphi_t'(y)B(y)\varphi_t'(y)^T - B\big(\varphi_t(y)\big)$. Using the variational equation for (4.1) prove that $\delta(t)$ is the solution of a homogeneous linear differential equation. Therefore, $\delta(0) = 0$ implies $\delta(t) = 0$ for all t.
7. Let $z = \vartheta(y)$ be a transformation taking the Poisson system $\dot y = B(y)\nabla H(y)$ to $\dot z = \widetilde B(z)\nabla K(z)$. Prove that $\Phi_h(y)$ is a Poisson integrator for $B(y)$ if and only if $\Psi_h(z) = \vartheta \circ \Phi_h \circ \vartheta^{-1}(z)$ is a Poisson integrator for $\widetilde B(z)$.
8. Let B be a skew-symmetric but otherwise arbitrary constant matrix, and consider the Poisson system $\dot y = B\nabla H(y)$. Prove that every symplectic Runge–Kutta method is a Poisson integrator for such a system.
   Hint. Transform B to block-diagonal form.
9. (M.J. Gander 1994). Consider the Lotka–Volterra equation (2.13) with separable Hamiltonian $H(u, v) = K(u) + L(v)$. Prove that

   $$u_{n+1} = u_n + hu_nv_nH_v(u_n, v_n), \quad v_{n+1} = v_n - hu_{n+1}v_nH_u(u_{n+1}, v_n)$$

   is a Poisson integrator for this system.
10. Find a change of coordinates that transforms the Lotka–Volterra system (2.14) into a Hamiltonian system (in canonical form). Following the approach of Example 4.11 construct Poisson integrators for this system.
11. Prove that the matrix $B(y)$ of Example 2.7 defines a Poisson bracket, by showing that the bracket is given as Dirac's bracket (Dirac 1950)

    $$\{F,G\} \;=\; \{\widehat F, \widehat G\} - \textstyle\sum_{i,j}\{\widehat F, c_i\}\gamma_{ij}\{c_j, \widehat G\}. \tag{7.2}$$

    Here F and G are functions of y, $\widehat F$ and $\widehat G$ are smooth functions of x satisfying $\widehat F(\chi(y)) = F(y)$ and $\widehat G(\chi(y)) = G(y)$, $c_i(x)$ are the constraint functions defining the manifold $\mathcal M$, and γ_{ij} are the entries of the inverse of the matrix $(\{c_i, c_j\})$. The Poisson bracket to the left in (7.2) corresponds to $B(y)$ and those to the right are the canonical brackets evaluated at $x = \chi(y)$. Replacing $\widehat F(x)$ by $\widehat F(x) + \sum_k \mu_k(x)c_k(x)$ with $\mu_k(x)$ such that $\{\widehat F, c_k\} = 0$ on $\mathcal M$ eliminates the sum in (7.2) and proves the Jacobi identity for $B(y)$.

Chapter VIII. Structure-Preserving Implementation

This chapter is devoted to practical aspects of an implementation of geometric integrators. We explain strategies for changing the step size which do not deteriorate the correct qualitative behaviour of the solution. We study multiple time stepping strategies, the effect of round-off in long-time integrations, and the efficient solution of nonlinear systems arising in implicit integration schemes.

VIII.1 Dangers of Using Standard Step Size Control

> Another possible shortcoming of the method concerns its behavior when used with a variable step size . . . The integrator completely loses its desirable qualities . . . This can be understood at least qualitatively by realizing that by changing the time step one is in essence continually changing the nearby Hamiltonian . . . (B. Gladman, M. Duncan & J. Candy 1991)

In the previous chapters we have studied symmetric and symplectic integrators, and we have seen an enormous progress in long-time integrations of various problems. Decades ago, a similar enormous progress was the introduction of algorithms with automatic step size control. Naively, one would expect that the blind combination of both techniques leads to even better performances. We shall see by a numerical experiment that this is not the case, a phenomenon observed by Gladman, Duncan & Candy (1991) and Calvo & Sanz-Serna (1992).

We study the long-time behaviour of symplectic methods combined with the following standard step size selection strategy (see e.g., Hairer, Nørsett & Wanner (1993), Sect. II.4). We assume that an expression err_n related to the local error is available for the current step computed with step size h_n (usually obtained with an embedded method). Based on an asymptotic formula $err_n \approx Ch_n^r$ (for $h_n \to 0$) and on the requirement to get an error close to a user supplied tolerance *Tol*, we predict a new step size by

$$h_{new} = 0.85 \cdot h_n \Big(\frac{Tol}{err_n} \Big)^{1/r}, \tag{1.1}$$

where a safety factor 0.85 is included. We then apply the method with step size $h_{n+1} = h_{new}$. If for the new step $err_{n+1} \leq Tol$, the step is accepted and the integration is continued. If $err_{n+1} > Tol$, it is rejected and recomputed with the step size h_{new} obtained from (1.1) with $n+1$ instead of n. Similar step size strategies are implemented in most codes for solving ordinary differential equations.

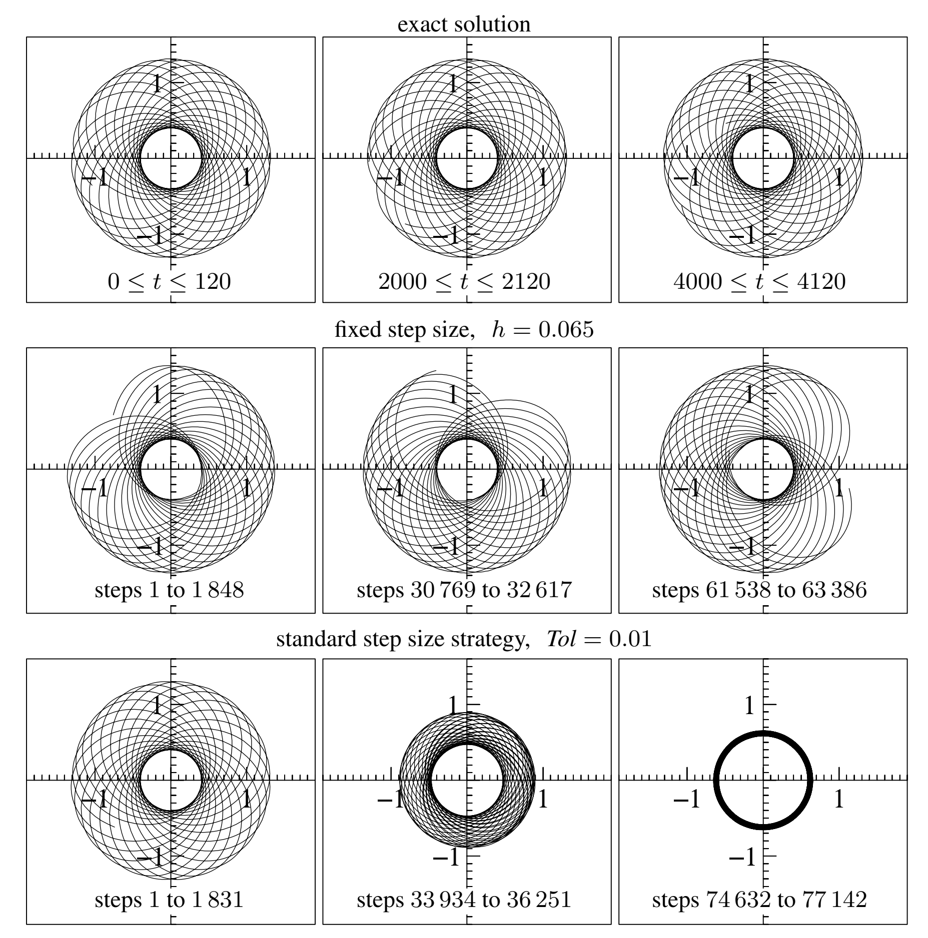


Fig. 1.1. Störmer–Verlet scheme applied with fixed step size (middle) or with the standard step size strategy (below) compared to the exact solution (above); solutions are for the interval $0 \le t \le 120$ (left), for $2000 \le t \le 2120$ (middle), and for $4000 \le t \le 4120$ (right)

Numerical Experiment. We consider the perturbed Kepler problem

$$\begin{aligned} \dot q_1 &= p_1, \qquad & \dot p_1 &= -\frac{q_1}{(q_1^2+q_2^2)^{3/2}} - \frac{\delta q_1}{(q_1^2+q_2^2)^{5/2}} \\ \dot q_2 &= p_2, \qquad & \dot p_2 &= -\frac{q_2}{(q_1^2+q_2^2)^{3/2}} - \frac{\delta q_2}{(q_1^2+q_2^2)^{5/2}} \end{aligned} \tag{1.2}$$

($\delta = 0.015$) with initial values

$$q_1(0) = 1-e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{(1+e)/(1-e)}$$

(eccentricity $e = 0.6$). As a numerical method we take the *Störmer–Verlet scheme* (I.1.17) which is symmetric, symplectic, and of order 2. The fixed step size imple-

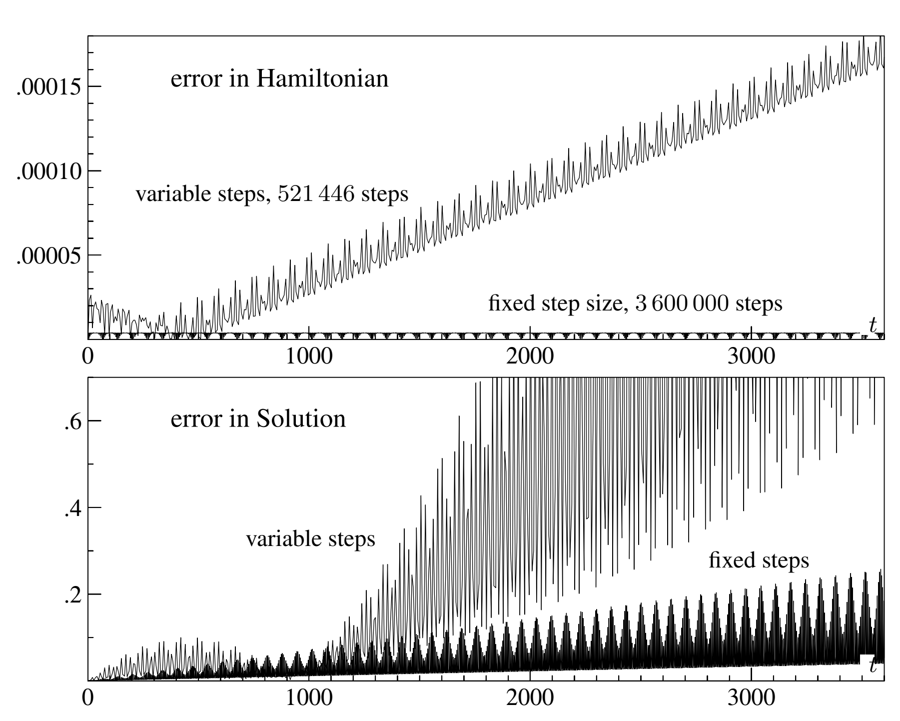


Fig. 1.2. Study of the error in the Hamiltonian and of the global error for the Störmer–Verlet scheme. Fixed step size implementation with $h = 10^{-3}$, variable step size with $Tol = 10^{-4}$

mentation is straightforward. For the variable step size strategy we take for err_n the Euclidean norm of the difference between the Störmer–Verlet solution and the symplectic Euler solution (which is available without any further function evaluation). Since $err_n = \mathcal{O}(h_n^2)$, we take $r = 2$ in (1.1).

The numerical solution in the (q_1, q_2)-plane is presented in Fig. 1.1. To make the long-time behaviour of the two implementations visible, we show the numerical solution on three different parts of the integration interval. We have included the numbers of steps needed for the integration to reach $t = 120$, 2120, and 4120, respectively. We see that the qualitative behaviour of the variable step size implementation is not correct, although it is more precise on short intervals. Moreover, the near-preservation of the Hamiltonian is lost (see Fig. 1.2) as is the linear error growth. Apparently, the error in the Hamiltonian behaves like $|a - bt|$ for the variable step size implementation, and that for the solution like $|ct - dt^2|$ (with constants a, b, c, d depending on *Tol*). Due to the relatively large eccentricity of the problem, the variable step size implementation needs fewer function evaluations for a given accuracy on a short time interval, but the opposite is true for long-time integrations.

The aim of the next two sections is to present approaches which permit the use of variable step sizes for symmetric or symplectic methods without losing the qualitatively correct long-time behaviour.

VIII.2 Time Transformations

A variable step size implementation produces approximations y_n on a (non-equidistant) grid $\{t_n\}$. The same effect can be achieved by performing in advance a time transformation $t \leftrightarrow \tau$ and by applying a constant step size implementation to the transformed system. If the time transformation is given as the solution of a differential equation, it follows from the chain rule $\frac{dy}{d\tau} = \frac{dy}{dt}\frac{dt}{d\tau}$ that the transformed system is

$$y' = \sigma(y)f(y), \qquad t' = \sigma(y). \tag{2.1}$$

Here, prime indicates a derivative with respect to τ, and we use the same letter y for the solutions $y(t)$ of $\dot y = f(y)$ and $y(\tau)$ of (2.1). If $\sigma(y) > 0$, the correspondence $t \leftrightarrow \tau$ is bijective.

Applying a numerical method with constant step size ε to (2.1) yields approximations $y_n \approx y(\tau_n) = y(t_n)$, where $\tau_n = n\varepsilon$ and

$$t_{n+1} - t_n = \int_{n\varepsilon}^{(n+1)\varepsilon} \sigma\big(y(\tau)\big)\, d\tau \approx \varepsilon\sigma(y_n). \tag{2.2}$$

Approximations to t_n are obtained by integrating numerically the differential equation $t' = \sigma(y)$ together with $y' = \sigma(y)f(y)$.

In the context of geometric numerical integration, we are interested in time transformations such that the vector field $\sigma(y)f(y)$ retains geometric features of $f(y)$.

VIII.2.1 Symplectic Integration

For a Hamiltonian system $\dot y = f(y) = J^{-1}\nabla H(y)$ it is natural to search for step size functions $\sigma(y)$ such that (2.1) is again Hamiltonian. For this we have to check whether the Jacobian of $\sigma(y)\nabla H(y)$ is symmetric (cf. Integrability Lemma VI.2.7). But this is the case only if $\nabla H(y)\nabla\sigma(y)^T$ is symmetric, i.e., $\nabla H(y)$ and $\nabla\sigma(y)$ are collinear, so that $\frac{d}{dt}\sigma\big(y(t)\big) = \nabla\sigma\big(y(t)\big)^T J\nabla H\big(y(t)\big) = 0$. Consequently, $\sigma(y) = Const$ along solutions of the Hamiltonian system which is what makes this approach unattractive for a variable step size integration. This disappointing fact has been observed by Stoffer (1988, 1995) and Skeel & Gear (1992).

The main idea for circumventing this difficulty is the following: suppose we want to integrate the Hamiltonian system with steps of size $h \approx \varepsilon\,\sigma(y)$, where $\sigma(y) > 0$ is a state-dependent given function and $\varepsilon > 0$ is a small parameter. Instead of multiplying the vector field $f(y) = J^{-1}\nabla H(y)$ by $\sigma(y)$, we consider the *new Hamiltonian*

$$K(y) = \sigma(y)\Big(H(y) - H_0\Big), \tag{2.3}$$

where $H_0 = H(y_0)$ for a fixed initial value y_0. The corresponding Hamiltonian system is

$$y' = \sigma(y)J^{-1}\nabla H(y) + \big(H(y) - H_0\big)J^{-1}\nabla\sigma(y). \tag{2.4}$$

Compared to (2.1) we have introduced a perturbation, which vanishes along the solution of the Hamiltonian system passing through y_0, but which makes the system Hamiltonian.

Time transformations such as in (2.3) are used in classical mechanics for an analytic treatment of Hamiltonian systems (Levi-Civita (1906, 1920), where (2.3) is called the "Darboux–Sundman transformation", see Sundman (1912)). Zare & Szebehely (1975) consider such time transformations for numerical purposes (without taking care of symplecticity). Waldvogel & Spirig (1995) apply the transformations proposed by Levi-Civita to Hill's lunar problem and solve the transformed equations by composition methods in order to preserve the symplectic structure. The following general procedure was proposed independently by Hairer (1997) and Reich (1999).

Algorithm 2.1. *Apply an arbitrary symplectic one-step method with constant step size ε to the Hamiltonian system (2.4), augmented by $t' = \sigma(y)$. This yields numerical approximations (y_n, t_n) with $y_n \approx y(t_n)$.*

Although this algorithm yields numerical approximations on a non-equidistant grid, it can be considered as a fixed step size, symplectic method applied to a different Hamiltonian system. This interpretation allows one to apply the standard techniques for the study of its long-time behaviour.

A disadvantage of this algorithm is that for separable Hamiltonians $H(p,q) = T(p)+U(q)$ the transformed Hamiltonian (2.3) is no longer separable. Hence, methods that are explicit for separable Hamiltonians are not explicit in the implementation of Algorithm 2.1. The following examples illustrate that this disadvantage can be partially overcome for the important case of Hamiltonian functions

$$H(p,q) = \frac{1}{2}\, p^T M^{-1} p + U(q), \tag{2.5}$$

where M is a constant symmetric matrix.

Example 2.2 (Symplectic Euler with *p*-Independent Step Size Function). For step size functions $\sigma(q)$ the symplectic Euler method, applied with constant step size ε to (2.4), reads

$$\begin{aligned} p_{n+1} &= p_n - \varepsilon\sigma(q_n)\nabla U(q_n) - \varepsilon\Big(\frac{1}{2}\, p_{n+1}^T M^{-1} p_{n+1} + U(q_n) - H_0\Big)\nabla\sigma(q_n) \\ q_{n+1} &= q_n + \varepsilon\sigma(q_n) M^{-1} p_{n+1} \end{aligned}$$

and yields an approximation at $t_{n+1} = t_n + \varepsilon\sigma(q_n)$. The first equation is nonlinear (quadratic) in p_{n+1}. Introducing the scalar quantity $\beta := \|p_{n+1}\|_M^2 := p_{n+1}^T M^{-1} p_{n+1}$, it reduces to the scalar quadratic equation

$$\beta = \Big\| p_n - \varepsilon\sigma(q_n)\nabla U(q_n) - \varepsilon\Big(\frac{\beta}{2} + U(q_n) - H_0\Big)\nabla\sigma(q_n)\Big\|_M^2$$

which can be solved directly. The numerical solution (p_{n+1}, q_{n+1}) is then given explicitly.

Choices of Step Size Functions. Sometimes suitable functions $\sigma(p,q)$ are known a priori. For example, for the two-body problem one can take $\sigma(p,q) = \|q\|^\alpha$, e.g., $\alpha = 2$, or $\alpha = 3/2$ to preserve the scaling invariance (Budd & Piggott 2003), so that smaller step sizes are taken when the two bodies are close.

An interesting choice, which does not require any a priori knowledge of the solution, is $\sigma(y) = \|f(y)\|^{-1}$. The solution of (2.1) then satisfies $\|y'(\tau)\| = 1$ (arc-length parameterization) and we get approximations y_n that are nearly equidistant in the phase space. Such time transformations have been proposed by McLeod & Sanz-Serna (1982) for graphical reasons and by Huang & Leimkuhler (1997). For a Hamiltonian system with $H(p,q)$ given by (2.5), it is thus natural to consider

$$\sigma(p,q) = \Big(\frac{1}{2}\,p^T M^{-1} p + \nabla U(q)^T M^{-1} \nabla U(q)\Big)^{-1/2}. \tag{2.6}$$

We have chosen this particular norm, because it leaves the expression (2.6) invariant with respect to linear coordinate changes $q \mapsto Aq$ (implying $p \mapsto A^{-T}p$). Exploiting the fact that the Hamiltonian (2.5) is constant along solutions, the step size function (2.6) can be replaced by the p-independent function

$$\sigma(q) = \Big(\big(H_0 - U(q)\big) + \nabla U(q)^T M^{-1} \nabla U(q)\Big)^{-1/2}. \tag{2.7}$$

The use of (2.6) and (2.7) gives nearly identical results, but (2.7) is easier to implement. If we are interested in an output that is approximatively equidistant in the q-space, we can take

$$\sigma(q) = \big(H_0 - U(q)\big)^{-1/2}. \tag{2.8}$$

Example 2.3 (Störmer–Verlet Scheme with *p*-Independent Step Size Function). For a step size function $\sigma(q)$ the Störmer–Verlet scheme gives

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{\varepsilon}{2}\sigma(q_n)\nabla U(q_n) - \frac{\varepsilon}{2}\Big(H\big(p_{n+1/2}, q_n\big) - H_0\Big)\nabla\sigma(q_n) \\ q_{n+1} &= q_n + \frac{\varepsilon}{2}\big(\sigma(q_n) + \sigma(q_{n+1})\big)M^{-1}p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{\varepsilon}{2}\sigma(q_{n+1})\nabla U(q_{n+1}) \\ &\qquad - \frac{\varepsilon}{2}\Big(H\big(p_{n+1/2}, q_{n+1}\big) - H_0\Big)\nabla\sigma(q_{n+1}). \end{aligned} \tag{2.9}$$

The first equation is essentially the same as that for the symplectic Euler method, and it can be solved for $p_{n+1/2}$ as explained in Example 2.2. The second equation is implicit in q_{n+1}, but it is sufficient to solve the scalar equation

$$\gamma = \sigma\Big(q_n + \frac{\varepsilon}{2}\big(\sigma(q_n) + \gamma\big)M^{-1}p_{n+1/2}\Big) \tag{2.10}$$

for $\gamma = \sigma(q_{n+1})$. Newton iterations can be efficiently applied, because $\nabla\sigma(q)$ is available already. The last equation (for p_{n+1}) is explicit. This variable step size Störmer–Verlet scheme gives approximations at t_n, where

$$t_{n+1} = t_n + \frac{\varepsilon}{2}\big(\sigma(q_n) + \sigma(q_{n+1})\big).$$

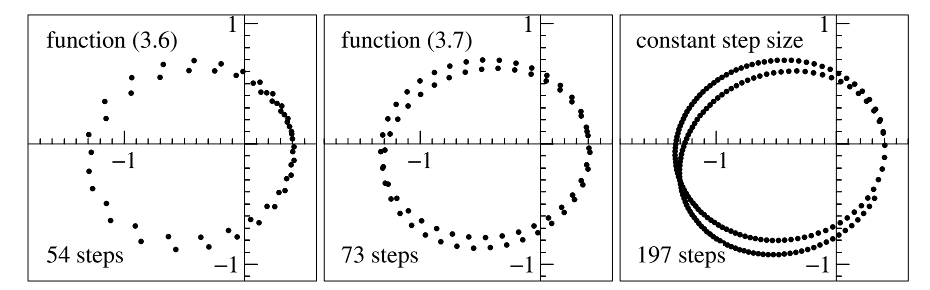


Fig. 2.1. Various step size strategies for the Störmer–Verlet scheme (Example 2.3) applied to the perturbed Kepler problem (1.2) on the interval $[0, 10]$ (approximately two periods)

In Fig. 2.1 we illustrate how the different step size functions influence the position of the output points. We apply the Störmer–Verlet method of Example 2.3 to the perturbed Kepler problem (1.2) with initial values, perturbation parameter, and eccentricity as in Sect. VIII.1. As step size functions we use (2.7), (2.8), and constant step size $\sigma(q) \equiv 1$. For all three choices of $\sigma(q)$ we have adjusted the parameter ε in such a way that the maximal error in the Hamiltonian is close to 0.01. The step size strategy (2.7) is apparently the most efficient one. For this strategy, we observe that the output points in the q-plane concentrate in regions where the velocity is large, while the constant step size implementation shows the opposite behaviour.

VIII.2.2 Reversible Integration

For ρ-reversible differential equations $\dot{y} = f(y)$, i.e., $f(\rho y) = -\rho f(y)$ for all y, the time transformed problem (2.1) remains ρ-reversible if

$$\sigma(\rho y) = \sigma(y). \tag{2.11}$$

This condition is not very restrictive and is satisfied by many important time transformations. In particular, (2.11) holds for the arc length parameterization $\sigma(y) = \|f(y)\|^{-1}$ if ρ is orthogonal. Consequently, it makes sense to apply symmetric, reversible numerical methods with constant step size ε directly to the system (2.1).

However, similar to the symplectic integration of Sect. VIII.2.1, there is a serious disadvantage. For separable differential equations (i.e., problems that can be split as $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$) and for non-constant $\sigma(p, q)$ the transformed system (2.1) is no longer separable. Hence, methods that are explicit for separable problems are not necessarily explicit for (2.1).

Example 2.4 (Adaptive Störmer–Verlet Method). We consider a Hamiltonian system with separable Hamiltonian (2.5), and we apply the Störmer–Verlet scheme to (2.1). This yields (Huang & Leimkuhler 1997)

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{\varepsilon}{2} s_n \nabla U(q_n) \\ q_{n+1} &= q_n + \frac{\varepsilon}{2}\big(s_n + s_{n+1}\big) M^{-1} p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{\varepsilon}{2} s_{n+1} \nabla U(q_{n+1}), \end{aligned} \tag{2.12}$$

where $s_n = \sigma(p_{n+1/2}, q_n)$ and $s_{n+1} = \sigma(p_{n+1/2}, q_{n+1})$ (notice that the s_{n+1} of the current step is not the same as the s_n of the subsequent step, if $\sigma(p,q)$ depends on p). The values (p_{n+1}, q_{n+1}) are approximations to the solution at t_n, where

$$t_{n+1} = t_n + \frac{\varepsilon}{2}\big(s_n + s_{n+1}\big).$$

For a p-independent step size function s, method (2.12) corresponds to that of Example 2.3, where the terms involving $\nabla\sigma(q)$ are removed. The implicitness of (2.12) is comparable to that of the method of Example 2.3. Completely explicit variants of this method will be discussed in the next section.

We conclude this section with a brief comparison of the variable step size Störmer–Verlet methods of Examples 2.3 and 2.4. Method (2.12) is easier to implement and more efficient when the step size function $\sigma(p,q)$ is expensive to evaluate. In a few numerical comparisons we observed, however, that the error in the Hamiltonian and in the solution is in general larger for method (2.12), and that the method (2.9) becomes competitive when $\sigma(p,q)$ is p-independent and easy to evaluate. A similar observation in favour of method (2.9) has been made by Calvo, López-Marcos & Sanz-Serna (1998).

VIII.3 Structure-Preserving Step Size Control

The disappointing long-time behaviour in Fig. 1.1 of the variable step size implementation of the Störmer–Verlet scheme is due to lack of reversibility. Indeed, for a ρ-reversible differential equation the step size $h_{n+1/2}$ taken for stepping from y_n to y_{n+1} should be the same as that when stepping from ρy_{n+1} to ρy_n (cf. Fig. V.1.1). The strategy of Sect. VIII.1, for which the step size depends on information of the preceding step, cannot guarantee such a property.

VIII.3.1 Proportional, Reversible Controllers

Following a suggestion of Stoffer (1988) we consider step sizes depending only on information of the present step, i.e., being *proportional* to some function of the actual state. This leads to the algorithm

$$y_{n+1} = \Phi_{h_{n+1/2}}(y_n), \qquad h_{n+1/2} = \varepsilon\, s(y_n, \varepsilon), \tag{3.1}$$

where $\Phi_h(y)$ is a one-step method for $\dot y = f(y)$, and ε is a small parameter. For theoretical investigations it is useful to consider the mapping

$$\Psi_\varepsilon(y) := \Phi_{\varepsilon s(y,\varepsilon)}(y). \tag{3.2}$$

This is a one-step discretization, consistent with $y' = s(y,0)f(y)$, and applied with constant step size ε. Consequently, all results concerning the long-time integration

with constant steps (e.g., backward error analysis of Chap. IX), and the definitions of symmetry and reversibility can be extended in a straightforward way.

Symmetry. We call the algorithm (3.1) symmetric, if $\Psi_\varepsilon(y)$ is symmetric, i.e., $\Psi_\varepsilon = \Psi_{-\varepsilon}^{-1}$. In the case of a symmetric Φ_h this is equivalent to

$$s(\widehat{y}, -\varepsilon) = s(y, \varepsilon) \qquad \text{with} \qquad \widehat{y} = \Phi_{\varepsilon s(y,\varepsilon)}(y). \tag{3.3}$$

Reversibility. The algorithm (3.1) is called ρ-reversible if, when applied to a ρ-reversible differential equation, $\Psi_\varepsilon(y)$ is ρ-reversible, i.e., $\rho \circ \Psi_\varepsilon = \Psi_\varepsilon^{-1} \circ \rho$ (cf. Definition V.1.2). If the method Φ_h is ρ-reversible then this is equivalent to

$$s(\rho^{-1}\widehat{y}, \varepsilon) = s(y, \varepsilon) \qquad \text{with} \qquad \widehat{y} = \Phi_{\varepsilon s(y,\varepsilon)}(y). \tag{3.4}$$

Example 3.1. Aiming at step sizes $h \approx \varepsilon\sigma(y)$ (cf. (2.2)), Hut, Makino & McMillan (1995) propose the use of $s(y,\varepsilon) = \frac{1}{2}\big(\sigma(y) + \sigma(\widehat{y})\big)$ where, as in Sect. VIII.2, $\sigma(y)$ is some function that uses an a priori knowledge of the solution of the differential equation. Notice that, because of $\widehat{y} = \Phi_{\varepsilon s(y,\varepsilon)}(y)$, the value of $s(y,\varepsilon)$ is defined by an implicit relation. Condition (3.3) is satisfied whenever $\Phi_h(y)$ is symmetric, and (3.4) is satisfied whenever $\Phi_h(y)$ is ρ-reversible and $\sigma(\rho y) = \sigma(y)$ holds. For a proof of these statements one shows that $s(\widehat{y}, -\varepsilon)$ and $s(y,\varepsilon)$ (resp. $s(\rho^{-1}\widehat{y}, \varepsilon)$ and $s(y,\varepsilon)$) are solution of the same nonlinear equation.

How can we find suitable step size functions $s(y,\varepsilon)$ which satisfy all these properties, and which do not require any a priori knowledge of the solution? In a remarkable publication, Stoffer (1995) gives the key to the answer of this question. He simply proposes to choose the step size h in such a way that the local error estimate satisfies $err = Tol$ (in contrast to $err \le Tol$ for the standard strategy). Let us explain this idea in some more detail for Runge–Kutta methods.

Example 3.2 (Symmetric, Variable Step Size Runge–Kutta Methods). For the numerical solution of $\dot{y} = f(y)$ we consider Runge–Kutta methods

$$Y_i = y_n + h\sum_{j=1}^{s} a_{ij} f(Y_j), \qquad y_{n+1} = y_n + h\sum_{i=1}^{s} b_i f(Y_i), \tag{3.5}$$

with coefficients satisfying $a_{s+1-i,s+1-j} + a_{ij} = b_j$ for all i, j. Such methods are symmetric and reversible (cf. Theorem V.2.3). A common approach for step size control is to consider an embedded method $\widehat{y}_{n+1} = y_n + h\sum_{i=1}^{s}\widehat{b}_i f(Y_i)$ (which has the same internal stages Y_i) and to take the difference $y_{n+1} - \widehat{y}_{n+1}$, i.e.,

$$D(y_n, h) = h\sum_{i=1}^{s} e_i f(Y_i) \tag{3.6}$$

with $e_i = b_i - \widehat{b}_i$, as indicator of the local error. For methods where $Y_i \approx y(t_n + c_i h)$ (e.g., collocation or discontinuous collocation) one usually computes the coefficients e_i from a nontrivial solution of the homogeneous linear system

$$\sum_{i=1}^{s} e_i c_i^{k-1} = 0 \qquad \text{for} \quad k = 1, \ldots, s-1. \tag{3.7}$$

This yields $D(y_n, h) = \mathcal{O}(h^r)$ with r close to s. According to the suggestion of Stoffer (1995) we determine the step size $h_{n+1/2}$ such that

$$\|D(y_n, h_{n+1/2})\| = \mathit{Tol}\,. \tag{3.8}$$

A Taylor expansion around $h = 0$ shows that $D(y, h) = d_r(y)h^r + \mathcal{O}(h^{r+1})$ with some $r \geq 1$. We assume $\|d_r(y)\| \neq 0$ and we put $\varepsilon = \mathit{Tol}^{1/r}$, so that $h_{n+1/2}$ from (3.8) can be expressed by a smooth function $s(y, \varepsilon)$ as (3.1).

To satisfy the *symmetry* relation (3.3) we determine the e_i such that

$$e_{s+1-i} = e_i \quad \text{for all } i \qquad \text{or} \qquad e_{s+1-i} = -e_i \quad \text{for all } i \tag{3.9}$$

(Hairer & Stoffer 1997). If the Runge–Kutta method is symmetric, this then implies

$$\|D(y_n, h)\| = \|D(y_{n+1}, -h)\| \qquad \text{with} \qquad y_{n+1} = \Phi_h(y_n). \tag{3.10}$$

This follows from the fact that the internal stage vectors Y_i of the step from y_n to y_{n+1} and the stage vectors $\overline{Y}_i$ of the step from y_{n+1} to y_n (negative step size $-h$) are related by $\overline{Y}_i = Y_{s+1-i}$. The step size determined by (3.8) is thus the same for both steps and, consequently, condition (3.3) holds.

The *reversibility* requirement (3.4) is a consequence of

$$\|D(y_n, h)\| = \|D(\rho^{-1}y_{n+1}, h)\| \qquad \text{with} \qquad y_{n+1} = \Phi_h(y_n) \tag{3.11}$$

which is satisfied for orthogonal mappings ρ (i.e., $\rho^T \rho = I$). This is seen as follows: applying Φ_h to $\rho^{-1}y_{n+1}$ gives $\rho^{-1}y_n$, and the internal stages are $\overline{Y}_i = \rho^{-1}Y_{s+1-i}$. Hence, we have from (3.9) that $D(\rho^{-1}y_{n+1}, h) = \pm\rho^{-1}D(y_n, h)$, and (3.11) follows from the orthogonality of ρ.

A simple special case is the trapezoidal rule

$$y_{n+1} = y_n + \frac{h_{n+1/2}}{2}\Big(f(y_n) + f(y_{n+1})\Big) \tag{3.12}$$

combined with

$$D(y_n, h) = \frac{h}{2}\Big(f(y_{n+1}) - f(y_n)\Big).$$

The scalar nonlinear equation (3.8) for $h_{n+1/2}$ can be solved in tandem with the nonlinear system (3.12).

Example 3.3 (Symmetric, Variable Step Size Störmer–Verlet Scheme). The strategy of Example 3.2 can be extended in a straightforward way to partitioned Runge–Kutta methods. For example, for the second order symmetric Störmer–Verlet scheme (I.1.17), applied to the problem $\dot q = p,\ \dot p = -\nabla U(q)$, we can take

$$D(p_n, q_n, h) = \frac{h}{2}\begin{pmatrix} \nabla U(q_{n+1}) - \nabla U(q_n) \\ h\big(\nabla U(q_{n+1}) + \nabla U(q_n)\big) \end{pmatrix}$$

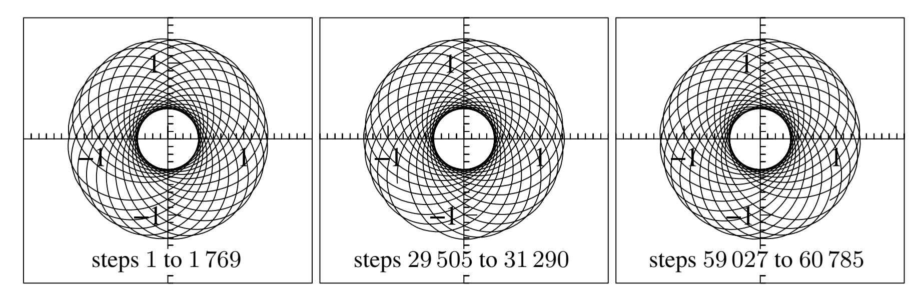


Fig. 3.1. Störmer–Verlet scheme applied with the symmetric adaptive step size strategy of Example 3.3 (*Tol* $= 0.01$); the three pictures have the same meaning as in Fig. 1.1

as error indicator. The first component is just the difference of the Störmer–Verlet solution and the numerical approximation obtained by the symplectic Euler method. The second component is a symmetrized version of it.

We apply this method with $h_{n+1/2}$ determined by (3.8) and *Tol* $= 0.01$ to the perturbed Kepler problem (1.2) with initial values as in Fig. 1.1. The result is given in Fig. 3.1. We identify a correct qualitative behaviour (compared to the wrong behaviour for the standard step size strategy in Fig. 1.1). It should be mentioned that the work for solving the scalar equation (3.8) for $h_{n+1/2}$ is not negligible, because the Störmer–Verlet scheme is explicit. Solving this equation iteratively, every iteration requires one force evaluation $\nabla U(q)$. An efficient solver for this scalar nonlinear equation should be used.

A Two-Step Proportional Controller. With the aim of obtaining a completely explicit integrator, Huang & Leimkuhler (1997) propose the use of two-term recurrence relations for the step size sequence, see also Holder, Leimkuhler & Reich (2001). Instead of using a relation between $h_{n+1/2}$, y_n and y_{n+1} (cf. Example 3.1) which is necessarily implicit, it is suggested to use a symmetric relation between $h_{n-1/2}$, $h_{n+1/2}$, and y_n, which then is explicit. In particular, with the notation $h_{n+1/2} = \varepsilon s_{n+1/2}$, it is proposed to use the two-term recurrence relation

$$\frac{1}{s_{n+1/2}} + \frac{1}{s_{n-1/2}} = \frac{2}{\sigma(y_n)}, \tag{3.13}$$

starting with $s_{1/2} = \sigma(y_0)$. In combination with the Störmer–Verlet method for separable Hamiltonians, this algorithm is completely explicit, and the authors report an excellent performance for realistic problems.

A rigorous analysis of the long-time behaviour of this variable step size Störmer–Verlet method is much more difficult. The results of Chapters IX and XI cannot be applied, because it is not a one-step mapping $y_n \mapsto y_{n+1}$. The analysis of Cirilli, Hairer & Leimkuhler (1999) shows that, similar to weakly stable multistep methods (Chap. XV), the numerical solution and the step size sequence contain oscillatory terms. Although these oscillations are usually very small (and hardly visible), it seems difficult to get rigorous estimates for them.

VIII.3.2 Integrating, Reversible Controllers

All variable step size approaches of this chapter are based on some time transformation $t \leftrightarrow \tau$ given by $\frac{dt}{d\tau} = \sigma(y)$ so that the differential equation, expressed in the new time variable τ, becomes

$$y' = \frac{1}{z}\, f(y), \qquad z\,\sigma(y) = 1. \tag{3.14}$$

In Sect. VIII.2 we insert $z^{-1} = \sigma(y)$ into the differential equation and apply a numerical method to $y' = \sigma(y)f(y)$. In Sect. VIII.3.1 we first discretize the algebraic relation $z\sigma(y) = 1$ expressing $z_{n+1/2}$ in terms of y_n and y_{n+1}, and then apply a one-step method to the differential equation in (3.14) assuming $z = z_{n+1/2}$ being constant.

In the present section we first differentiate the algebraic relation of (3.14) with respect to τ. This yields by Leibniz' rule $z'\sigma(y) + z\nabla\sigma(y)^T y' = 0$ so that

$$z' = G(y) \qquad \text{with} \qquad G(y) = -\frac{1}{\sigma(y)}\,\nabla\sigma(y)^T f(y). \tag{3.15}$$

The idea of differentiating the constraint in (3.14) has been raised in Huang & Leimkuhler (1997), but soon abandoned in favour of the controller (3.13). The subsequent algorithm together with its theoretical justification is elaborated in Hairer & Söderlind (2004). The idea is to discretize first the differential equation in (3.15) and then to apply a one-step method to the problem (3.14) with constant z. The proposed algorithm is thus

$$\begin{aligned} z_{n+1/2} &= z_{n-1/2} + \varepsilon\, G(y_n) \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \end{aligned} \tag{3.16}$$

with $z_{1/2} = z_0 + \varepsilon\, G(y_0)/2$ and $z_0 = 1/\sigma(y_0)$. This algorithm is explicit whenever the underlying one-step method $\Phi_h(y)$ is explicit. It is called *integrating* controller, because the step size density is obtained by summing up small quantities.

For a theoretical analysis it is convenient to introduce $z_n = (z_{n+1/2}+z_{n-1/2})/2$ and to write (3.16) as a one-step method for the augmented system

$$y' = \frac{1}{z}\, f(y), \qquad z' = G(y). \tag{3.17}$$

Notice that $I(y,z) = z\,\sigma(y)$ is a first integral of this system.

Algorithm 3.4. *Let $\Phi_h(y)$ be a one-step method for $\dot y = f(y)$, $y(0) = y_0$. With $G(y)$ given by (3.15), $z_0 = 1/\sigma(y_0)$, and constant ε, we let*

$$\begin{aligned} z_{n+1/2} &= z_n + \varepsilon\, G(y_n)/2 \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \\ z_{n+1} &= z_{n+1/2} + \varepsilon\, G(y_{n+1})/2. \end{aligned} \tag{3.18}$$

The values y_n approximate $y(t_n)$, where $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$.

This algorithm has an interesting interpretation as Strang splitting for the solution of (3.17): it approximates the flow of $z' = G(y)$ with fixed y over a half-step $\varepsilon/2$; then applies the method Φ_ε to $y' = f(y)/z$ with fixed z; finally, it computes a second half-step of $z' = G(y)$ with fixed y.

With the notation

$$\widehat{\Phi}_\varepsilon : \begin{pmatrix} y_n \\ z_n \end{pmatrix} \mapsto \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} \qquad \text{and} \qquad \widehat{\rho} = \begin{pmatrix} \rho & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.19}$$

the Algorithm 3.4 has the following properties:

- $\widehat{\Phi}_\varepsilon$ is symmetric whenever Φ_h is symmetric;
- $\widehat{\Phi}_\varepsilon$ is reversible with respect to $\widehat{\rho}$ whenever Φ_h is reversible with respect to ρ and $G(\rho y) = -G(y)$ (this is a consequence of $\sigma(\rho y) = \sigma(y)$).

These properties imply that standard techniques for constant step size implementations can be applied to $\widehat{\Phi}_\varepsilon$, and thus yield insight into the variable step size algorithm of this section. It will be shown in Chap. XI that when applied to integrable reversible systems there is no drift in the action variables and the global error grows only linearly with time. Moreover, the first integral $I(y,z) = z\,\sigma(y)$ of the system (3.17) is also well preserved (without drift) for such problems.

Example 3.5 (Variable Step Size Störmer–Verlet method). Consider a Hamiltonian system with separable Hamiltonian $H(p,q) = T(p) + U(q)$. Using the Störmer–Verlet method as basic method the above algorithm becomes (starting with $z_0 = 1/\sigma(y_0)$ and $z_{1/2} = z_0 + \varepsilon\, G(p_0, q_0)/2$)

$$\begin{aligned} z_{n+1/2} &= z_{n-1/2} + \varepsilon\, G(p_n, q_n) \\ p_{n+1/2} &= p_n - \varepsilon\, \nabla U(q_n)/(2z_{n+1/2}) \\ q_{n+1} &= q_n + \varepsilon\, \nabla T(p_{n+1/2})/z_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \varepsilon\, \nabla U(q_{n+1})/(2z_{n+1/2}). \end{aligned} \tag{3.20}$$

This method is explicit, symmetric and reversible as long as $G\rho = -G$, and computes approximations on a non-equidistant grid $\{t_n\}$ given by $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$.

Let us apply this method to the perturbed Kepler problem with data and initial values as in the beginning of this chapter. Further, we select $\sigma(q) = (q^T q)^{\alpha/2}$ with $\alpha = 3/2$, so that the control function (3.15) becomes

$$G(p,q) = -\alpha\, p^T q / q^T q\,. \tag{3.21}$$

Figure 3.2 shows the error in the Hamiltonian along the numerical solution as well as the global error in the solution (fictive step size $\varepsilon = 0.02$). The error in the Hamiltonian is proportional to ε^2 without drift, and the global error grows linearly with time (in double logarithmic scale a linear growth corresponds to a line with slope one; such lines are drawn in grey). This is qualitatively the same behaviour as observed in constant step size implementations of symplectic methods.

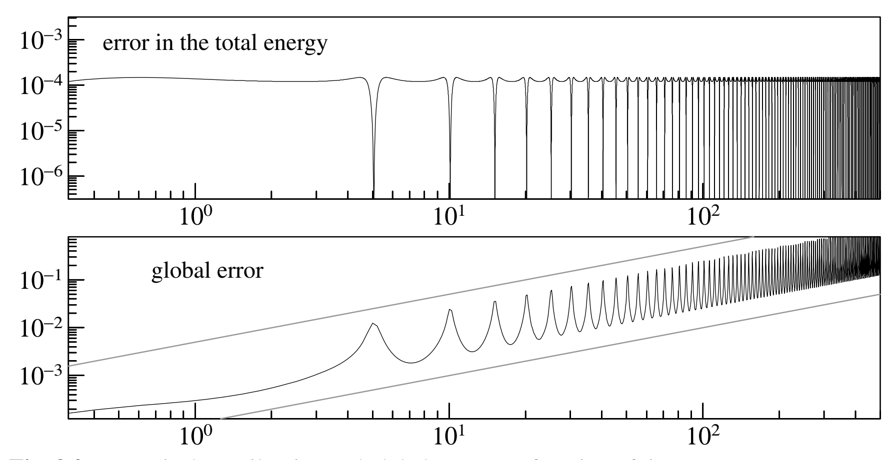


Fig. 3.2. Numerical Hamiltonian and global error as a function of time

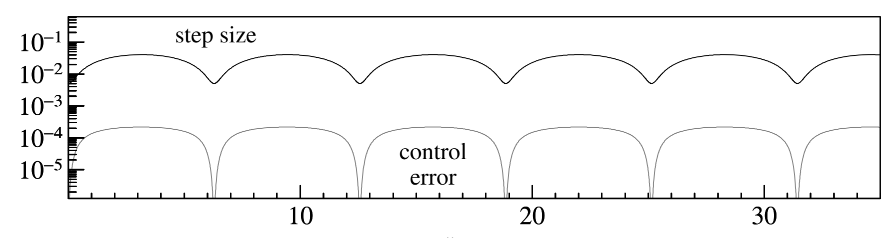


Fig. 3.3. Step sizes of the variable step size Störmer–Verlet method as a function of time, and the control error $z_n\sigma(q_n) - z_0\sigma(q_0)$ (grey curve)

Figure 3.3 shows the selected step sizes $h_{n+1/2} = \varepsilon/z_{n+1/2}$ as a function of time, and the control error $z_n\sigma(q_n) - z_0\sigma(q_0)$ in grey. Since its deviation from the constant value $z_0\sigma(q_0) = 1$ is small without any drift, the step density remains close to $1/\sigma(q)$. For an explanation of this excellent long-time behaviour we refer to Sect. XI.3.

VIII.4 Multiple Time Stepping

A completely different approach to variable step sizes will be described in this section. We are interested in situations where:

- many solution components of the differential equation vary slowly and only a few components have fast dynamics; or
- computationally expensive parts of the right-hand side do not contribute much to the dynamics of the solution.

In the first case it is tempting to use large step sizes for the slow components and small step sizes for the fast ones. Such integrators, called *multirate methods*, were

first formulated by Rice (1960) and Gear & Wells (1984). They were further developed by Günther & Rentrop (1993) in view of applications in electric circuit simulation, and by Engstler & Lubich (1997) with applications in astrophysics. Symmetric multirate methods are obtained from the approaches described below and are specially constructed by Leimkuhler & Reich (2001).

The second case suggests the use of methods that evaluate the expensive part of the vector field less often than the rest. This approach is called *multiple time stepping*. It was originally proposed for astronomy by Hayli (1967) and has become very popular in molecular dynamics simulations (Streett, Tildesley & Saville 1978, Grubmüller, Heller, Windemuth & Schulten 1991, Tuckerman, Berne & Martyna 1992). As noticed by Biesiadecki & Skeel (1993), one approach to such methods is within the framework of splitting and composition methods, which yields symmetric and symplectic methods. A second family of symmetric multiple time stepping methods results from the concept of using averaged force evaluations.

VIII.4.1 Fast-Slow Splitting: the Impulse Method

Consider a differential equation

$$\dot{y} = f(y), \qquad f(y) = f^{[\text{slow}]}(y) + f^{[\text{fast}]}(y), \tag{4.1}$$

where the vector field is split into summands contributing to slow and fast dynamics, respectively, and where $f^{[\text{slow}]}(y)$ is more expensive to evaluate than $f^{[\text{fast}]}(y)$. Multirate methods can often be cast into this framework by collecting in $f^{[\text{slow}]}(y)$ those components of $f(y)$ which produce slow dynamics and in $f^{[\text{fast}]}(y)$ the remaining components.

Algorithm 4.1. *For a given $N \geq 1$ and for the differential equation (4.1) a multiple time stepping method is obtained from*

$$\big(\Phi^{[\text{slow}]}_{h/2}\big)^* \circ \big(\Phi^{[\text{fast}]}_{h/N}\big)^N \circ \Phi^{[\text{slow}]}_{h/2}, \tag{4.2}$$

where $\Phi^{[\text{slow}]}_h$ and $\Phi^{[\text{fast}]}_h$ are numerical integrators consistent with $\dot{y} = f^{[\text{slow}]}(y)$ and $\dot{y} = f^{[\text{fast}]}(y)$, respectively.

The method of Algorithm 4.1 is already stated in symmetrized form (Φ^*_h denotes the adjoint of Φ_h). It is often called the *impulse method*, because the slow part $f^{[\text{slow}]}$ of the vector field is used – impulse-like – only at the beginning and at the end of the step, whereas the many small substeps in between are concerned solely through integrating the fast system $\dot{y} = f^{[\text{fast}]}(y)$.

Lemma 4.2. *Let $\Phi^{[\text{slow}]}_h$ be an arbitrary method of order 1, and $\Phi^{[\text{fast}]}_h$ a symmetric method of order 2. Then, the multiple time stepping algorithm (4.2) is symmetric and of order 2.*

If $f^{[\text{slow}]}(y)$ and $f^{[\text{fast}]}(y)$ are Hamiltonian and if $\Phi^{[\text{slow}]}_h$ and $\Phi^{[\text{fast}]}_h$ are both symplectic, then the multiple time stepping method is also symplectic.

Proof. Due to the interpretation of multiple time stepping as composition methods the proof of these statements is obvious. □

The order statement of Lemma 4.2 is valid for $h \to 0$, but should be taken with caution if the product of the step size h with a Lipschitz constant of the problem is not small (see Chap. XIII for a detailed analysis): it is *not* stated, and is not true in general for large N, that if h and h/N are the step sizes needed to integrate the slow and fast system, respectively, with an error bounded by ε, then the error of the combined scheme is $\mathcal{O}(\varepsilon)$.

The most important application of multiple time stepping is in Hamiltonian systems with a separable Hamiltonian

$$H(p,q) = T(p) + U(q), \qquad U(q) = U^{[\text{slow}]}(q) + U^{[\text{fast}]}(q). \tag{4.3}$$

If we let the fast vector field correspond to $T(p) + U^{[\text{fast}]}(q)$ and the slow vector field to $U^{[\text{slow}]}(q)$, and if we apply the Störmer–Verlet method and exact integration, respectively, Algorithm 4.1 reads

$$\varphi_{h/2}^{[\text{slow}]} \circ \big(\varphi_{h/2N}^{[\text{fast}]} \circ \varphi_{h/N}^{T} \circ \varphi_{h/2N}^{[\text{fast}]}\big)^N \circ \varphi_{h/2}^{[\text{slow}]}, \tag{4.4}$$

where $\varphi_t^T, \varphi_t^{[\text{slow}]}, \varphi_t^{[\text{fast}]}$ are the exact flows corresponding to the Hamiltonian systems for $T(p), U^{[\text{slow}]}(q), U^{[\text{fast}]}(q)$, respectively. Notice that for $N = 1$ the method (4.4) reduces to the Störmer–Verlet scheme applied to the Hamiltonian system with $H(p,q)$. This is a consequence of the fact that $\varphi_t^{[\text{fast}]} \circ \varphi_t^{[\text{slow}]} = \varphi_t^U$ is the exact flow of the Hamiltonian system corresponding to $U(q)$ of (4.3). In the molecular dynamics literature, the method (4.4) is known as the Verlet-I method (Grubmüller et al. 1991, who consider the method with little enthusiasm) or r-RESPA method (Tuckerman et al. 1992, with much more enthusiasm).

Example 4.3. In order to illustrate the effect of multiple time stepping we choose a 'solar system' with two planets, i.e., with a Hamiltonian

$$H(p,q) = \frac{1}{2}\Big(\frac{p_0^T p_0}{m_0} + \frac{p_1^T p_1}{m_1} + \frac{p_2^T p_2}{m_2}\Big) - \frac{m_0 m_1}{\|q_0 - q_1\|} - \frac{m_0 m_2}{\|q_0 - q_2\|} - \frac{m_1 m_2}{\|q_1 - q_2\|},$$

where $m_0 = 1, m_1 = m_2 = 10^{-2}$ and initial values $q_0 = (0,0)$, $\dot q_0 = (0,0)$, $q_1 = (1,0)$, $\dot q_1 = (0,1)$, $q_2 = (4,0)$, $\dot q_2 = (0,0.5)$. With these data, the motion of the two planets is nearly circular with periods close to 2π and 14π, respectively.

We split the potential as

$$U^{[\text{fast}]}(q) = -\frac{m_0 m_1}{\|q_0 - q_1\|}, \qquad U^{[\text{slow}]}(q) = -\frac{m_0 m_2}{\|q_0 - q_2\|} - \frac{m_1 m_2}{\|q_1 - q_2\|},$$

and we apply the algorithm of (4.4) with $N = 1$ (Störmer–Verlet), $N = 4$, and $N = 8$. Since the evaluation of $\varphi_t^{[\text{slow}]}$ is about twice as expensive as $\varphi_t^{[\text{fast}]}$ and that of φ_t^T is of negligible cost, the computational work of applying (4.4) on a fixed interval is proportional to

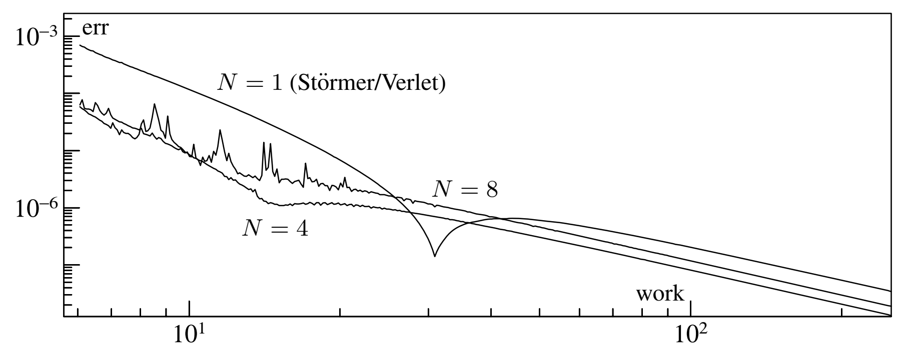


Fig. 4.1. Maximal error in the Hamiltonian as a function of computational work

$$\frac{2\pi}{h} \cdot \frac{(2+N)}{3}. \tag{4.5}$$

Our computations have shown that this measure of work corresponds very well to the actual cpu time.

We have solved this problem with many different step sizes h. Figure 4.1 shows the maximal error in the Hamiltonian (over the interval $[0, 200\pi]$) as a function of the computational work (4.5). We notice that the value $N = 4$ yields excellent results for relatively large as well as small step sizes. It noticeably improves the performance of the Störmer–Verlet method. If N becomes too large, an irregular behaviour for large step sizes is observed. Such “artificial resonances” are notorious for this method and have been discussed by Biesiadecki & Skeel (1993) for a similar experiment; also see Chap. XIII. For large N we also note a loss of accuracy for small step sizes. The optimal choice of N (which here is close to 4) depends on the problem and on the splitting into fast and slow parts, and has to be determined by experiment.

The multiple time stepping technique can be iteratively extended to problems with more than two different time scales. The idea is to split the ‘fast’ vector field of (4.1) into $f^{[\text{fast}]}(y) = f^{[ff]}(y) + f^{[fs]}(y)$, and to replace the method $\Phi_h^{[\text{fast}]}$ in Algorithm 4.1 with a multiple time stepping method. Depending on the problem, a significant gain in computer time may be achieved in this way.

Many more multiple time stepping methods that extend the above Verlet-I/r-RESPA/impulse method, have been proposed in the literature, most notably the mollified impulse method of García-Archilla, Sanz-Serna & Skeel (1999); see Sect. XIII.1.

VIII.4.2 Averaged Forces

A different approach to multiple time stepping arises from the idea of advancing the step with *averaged force evaluations*. We describe such a method for the second-order equation

$$\ddot{y} = f(y), \qquad f(y) = f^{[\text{slow}]}(y) + f^{[\text{fast}]}(y). \tag{4.6}$$

The exact solution satisfies

$$y(t+h) - 2y(t) + y(t-h) = h^2 \int_{-1}^{1} (1 - |\theta|)\, f\big(y(t+\theta h)\big)\, d\theta \;,$$

where the integral on the right-hand side represents a weighted average of the force along the solution, which is now going to be approximated. At $t = t_n$, we replace

$$f\big(y(t_n + \theta h)\big) \approx f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}\big(u(\theta h)\big)$$

where $u(\tau)$ is a solution of the differential equation

$$\ddot{u} = f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}(u) \;. \tag{4.7}$$

We then have

$$h^2 \int_{-1}^{1} (1 - |\theta|) \Big(f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}\big(u(\theta h)\big)\Big) d\theta = u(h) - 2u(0) + u(-h) \;.$$

The velocities are treated similarly, starting from the identity

$$\dot{y}(t+h) - \dot{y}(t-h) = h \int_{-1}^{1} f\big(y(t+\theta h)\big)\, d\theta \;.$$

A Symmetric Two-Step Method. For the differential equation (4.7) we assume the initial values

$$u(0) = y_n, \quad \dot{u}(0) = \dot{y}_n \;. \tag{4.8}$$

This initial value problem is solved numerically, e.g., by the Störmer–Verlet method with a smaller step size $\pm h/N$ on the interval $[-h, h]$, yielding numerical approximations $u_N(\pm h)$ and $v_N(\pm h)$ to $u(\pm h)$ and $\dot{u}(\pm h)$, respectively. Note that no further evaluations of $f^{[\text{slow}]}$ are needed for the computation of $u_N(\pm h)$ and $v_N(\pm h)$. This finally gives the symmetric two-step method (Hochbruck & Lubich 1999a)

$$\begin{aligned} y_{n+1} - 2y_n + y_{n-1} &= u_N(h) - 2u_N(0) + u_N(-h) \\ \dot{y}_{n+1} - \dot{y}_{n-1} &= v_N(h) - v_N(-h) \;. \end{aligned} \tag{4.9}$$

The starting values y_1 and $\dot{y}_1$ are chosen as $u_N(h)$ and $v_N(h)$ which correspond to (4.7) and (4.8) for $n = 0$.

A Symmetric One-step Method. An explicit one-step method with similar averaged forces is obtained when the initial values for (4.7) are chosen as

$$u(0) = y_n, \quad \dot{u}(0) = 0 \;. \tag{4.10}$$

It may appear crude to take zero initial values for the velocity, but we remark that for linear $f^{[\text{fast}]}$ the averaged force $(u(h) - 2u(0) + u(-h))/h^2$ does not depend on

the choice of $\dot{u}(0)$. Moreover the solution then satisfies $u(-t) = u(t)$, so that the computational cost is halved. We again denote by $u_N(h) = u_N(-h)$ the numerical approximation to $u(h)$ obtained with step size $\pm h/N$ from a one-step method (e.g., from the Störmer–Verlet scheme). Because of (4.10) the averaged forces

$$F_n = \frac{1}{h^2}\big(u_N(h) - 2u_N(0) + u_N(-h)\big) = \frac{2}{h^2}\big(u_N(h) - u_N(0)\big)$$

now depend only on y_n and not on the velocity $\dot{y}_n$. In trustworthy Verlet manner, the scheme $y_{n+1} - 2y_n + y_{n-1} = h^2 F_n$ can be written as the one-step method

$$\begin{aligned} v_{n+1/2} &= v_n + \frac{h}{2} F_n \\ y_{n+1} &= y_n + h v_{n+1/2} \\ v_{n+1} &= v_{n+1/2} + \frac{h}{2} F_{n+1} . \end{aligned} \tag{4.11}$$

The auxiliary variables v_n can be interpreted as averaged velocities: we have

$$v_n = \frac{y_{n+1} - y_{n-1}}{2h} \approx \frac{y(t_{n+1}) - y(t_{n-1})}{2h} = \frac{1}{2}\int_{-1}^{1} \dot{y}(t_n + \theta h)\, d\theta .$$

This average may differ substantially from $\dot{y}(t_n)$ if the solution is highly oscillatory in $[-h, h]$. In the experiments of this section it turned out that the choice $v_0 = \dot{y}_0$ and $\dot{y}_n = v_n$ as velocity approximations gives excellent results.

In a multirate context, symmetric one-step schemes using averaged forces were studied by Hochbruck & Lubich (1999b), Nettesheim & Reich (1999), and Leimkuhler & Reich (2001). A closely related approach for problems with multiple time scales is the heterogeneous multiscale method by E (2003) and Engquist & Tsai (2005).

Example 4.4. We add a satellite of mass $m_3 = 10^{-4}$ to the three body-problem of Example 4.3. It moves rapidly around the planet number one. The initial positions and velocities are $q_3 = (1.01, 0)$ and $p_3 = (0, 0)$. We split the potential as

$$U^{[\text{fast}]}(q) = -\frac{m_1 m_3}{\|q_1 - q_3\|}, \qquad U^{[\text{slow}]}(q) = -\sum_{\substack{i<j \\ (i,j)\neq(1,3)}} \frac{m_i m_j}{\|q_i - q_j\|},$$

and we apply the methods (4.9), (4.11), and the impulse method (4.4). Since the sum in $U^{[\text{slow}]}$ contains 5 terms, the computational work is proportional to

$$\begin{aligned} &\frac{5+N}{6h} \qquad \text{for methods (4.11) and (4.4)} \\ &\frac{6+2N}{6h} \qquad \text{for method (4.9).} \end{aligned}$$

For each of the methods we have optimized the number N of small steps. We obtained a flat minimum near $N = 40$ for (4.9) and (4.4), and a more pronounced minimum at $N = 12$ for (4.11). Figure 4.2 shows the errors at $t = 10$ in the positions and in the Hamiltonian as a function of the computational work.

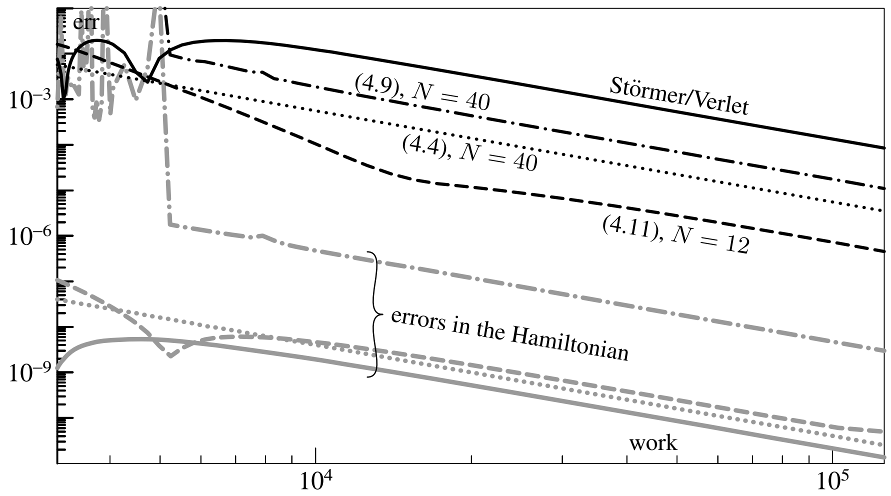


Fig. 4.2. Errors in position and in the Hamiltonian as a function of the computational work; the classical Störmer–Verlet method, the impulse method (4.4), and the averaged force methods (4.11) and (4.9). The errors in the Hamiltonian are indicated by grey lines (same linestyle)

The error in the position is largest for the Störmer–Verlet method and significantly smallest for the one-step averaged-force method (4.11). The errors in the velocities are about a factor 100 larger for all methods. They are not included in the figure. The error in the Hamiltonian is very similar for all methods with the exception of the two-step averaged-force method (4.9), for which it is much larger.

VIII.5 Reducing Rounding Errors

> . . . the idea is to capture the rounding errors and feed them back into the summation. (N.J. Higham 1993)

All numerical methods for solving ordinary differential equations require the computation of a recursion of the form

$$y_{n+1} = y_n + \delta_n, \tag{5.1}$$

where δ_n, the increment, is usually smaller in magnitude than the approximation y_n to the solution. In this situation the rounding errors caused by the computation of δ_n are in general smaller than those due to the addition in (5.1).

A first attempt at reducing the accumulation of rounding errors (in fixed-point arithmetic for his Runge–Kutta code) was due to Gill (1951). Kahan (1965) and Möller (1965) both extended this idea to floating point arithmetic. The resulting algorithm is nowadays called ‘compensated summation’, and a particularly nice presentation and analysis is given by N. Higham (1993). In the following algorithm we assume that y_n is a scalar; vector valued recursions are treated componentwise.

Algorithm 5.1 (Compensated Summation). *Let y_0 and $\{\delta_n\}_{n\geq 0}$ be given and put $e = 0$. Compute $y_1, y_2, \ldots$ from (5.1) as follows:*

for $n = 0, 1, 2, \ldots$ *do*
 $a = y_n$
 $e = e + \delta_n$
 $y_{n+1} = a + e$
 $e = e + (a - y_{n+1})$
end do

This algorithm can best be understood with the help of Fig. 5.1 (following the presentation of N. Higham (1993)). We present the mantissas of floating point numbers by boxes, for which the horizontal position indicates the exponent (for a large exponent the box is more to the left). The mantissas of y_n and e together represent the accurate value of y_n (notice that in the beginning $e = 0$). The operations of Algorithm 5.1 yield y_{n+1} and a new e, which together represent $y_{n+1} = y_n + \delta_n$. No digit of δ_n is lost in this way. With a standard summation the last digits of δ_n (those indicated by δ'' in Fig. 5.1) would have been missed.

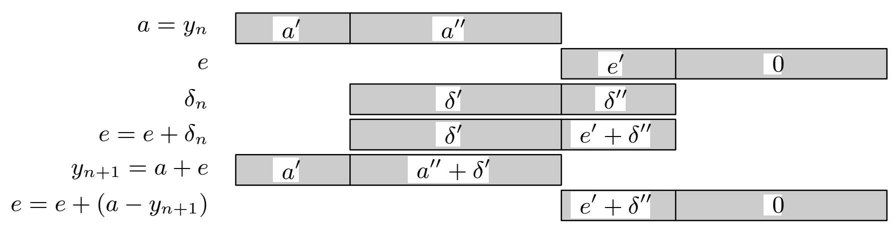


Fig. 5.1. Illustration of the technique of "compensated summation"

Numerical Experiment. We study the effect of compensated summation on the Kepler problem (I.2.2) (written as a first order system) with eccentricity $e = 0.6$ and initial values as in (I.2.11), so that the period of the elliptic orbit is exactly 2π. As the numerical integrator we take the composition method (V.3.13) of order 8 with the Störmer–Verlet scheme as basic integrator. We compute the numerical solution with step size $h = 2\pi/500$ once with standard update of the increment, once with compensated summation (both in double precision) and, in order to get a reference solution, we also perform the whole computation in quadruple precision. The difference between the double and quadruple precision computations gives us the rounding errors. Their Euclidean norms as a function of time are displayed in Fig. 5.2.

We see that throughout the whole integration interval the rounding errors of the standard implementation are nearly a factor of 100 larger than those of the implementation with compensated summation. This corresponds to the inverse of the step size or, more precisely, to the mean quotient between y_n and δ_n in (5.1). In Fig. 5.2 we have also included the pure global error of the method (without rounding errors) at integral multiples of the period 2π (hence no oscillations are visible). This is

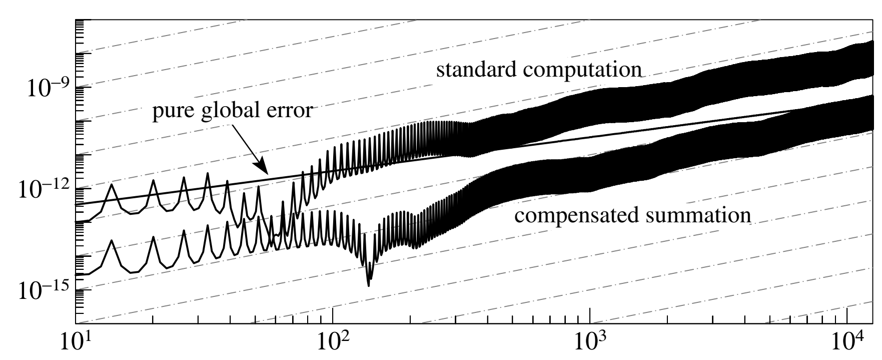


Fig. 5.2. Rounding errors and pure global error as a function of time; the parallel grey lines indicate a growth of $\mathcal{O}(t^{3/2})$

obtained as the difference of the numerical solution computed with quadruple precision and the exact solution. We observe a linear growth of the pure global error (this will be explained in Sect. X.3) and a growth like $\mathcal{O}(t^{3/2})$ due to the rounding errors. Thus, eventually the rounding errors will surpass the truncation errors, but this happens for the compensated summation only after some 1000 periods.

Probabilistic Explanation of the Error Growth. Our aim is to explain the growth rate of rounding errors observed in Fig. 5.2. Denote by ε_k the vector of rounding errors produced during the computations in the kth step. Since the derivative of the flow $\varphi_t(y)$ describes the propagation of these errors, the accumulated rounding error at time $t = t_N$ ($t_k = kh$) is

$$\eta_t = \sum_{k=1}^{N} \varphi'_{t-t_k}(y_k)\varepsilon_k. \tag{5.2}$$

For the Kepler problem and, in fact, for all completely integrable differential equations (cf. Sect. X.1) the flow and its derivative grow at most linearly with time, i.e.,

$$\big\|\varphi'_{t-t_k}(y)\big\| \le a + b(t - t_k) \qquad \text{for} \quad t \ge t_k. \tag{5.3}$$

Using $\varepsilon_k = \mathcal{O}(eps)$, where *eps* denotes the roundoff unit of the computer, an application of the triangle inequality to (5.2) yields $\eta_t = \mathcal{O}(t^2 eps)$. From our experiment of Fig. 5.2 we see that such an estimate is too pessimistic.

For a better understanding of accumulated rounding errors over long time intervals we make use of probability theory. Such an approach has been developed in the classical book of Henrici (1962). We assume that the components ε_{ki} of ε_k are *random variables* with mean and variance

$$E(\varepsilon_{ki}) = 0, \qquad \mathit{Var}\,(\varepsilon_{ki}) = C_{ki} \cdot \mathit{eps}^2,$$

and uniformly bounded $C_{ki} \le C$. For simplicity we assume that all ε_{ki} are independent random variables. Replacing the matrix $\varphi_{t-t_k}(y_k)$ in (5.2) with $\varphi_{t-t_k}\big(y(t_k)\big)$

and denoting its entries by w_{ijk}, the ith component of the accumulated rounding error (5.2) becomes

$$\eta_{ti} = \sum_{k=1}^{N} \sum_{j=1}^{n} w_{ijk} \varepsilon_{kj},$$

a linear combination of the random variables ε_{kj}. Elementary probability theory thus implies that

$$E(\eta_{ti}) = 0 \qquad \text{and} \qquad \mathit{Var}\,(\eta_{ti}) = \sum_{k=1}^{N} \sum_{j=1}^{n} w_{ijk}^2 \mathit{Var}\,(\varepsilon_{kj}).$$

Inserting the estimate (5.3) for w_{ijk} we get

$$\mathit{Var}\,(\eta_{ti}) \;\le\; \sum_{k=1}^{N} \big(a + b(t - t_k)\big)^2 \max_{j=1,\ldots,n} \mathit{Var}\,(\varepsilon_{kj}) \;=\; \mathcal{O}\Big(\frac{C}{h}\, t^3\, \mathit{eps}^2\Big).$$

Consequently, the Euclidean norm of the expected rounding error η_t is

$$\Big(\sum_{i=1}^{n} \mathit{Var}\,(\eta_{ti})\Big)^{1/2} \;=\; \mathcal{O}\Big(\sqrt{\frac{C}{h}}\, t^{3/2}\, \mathit{eps}\Big).$$

This is in excellent agreement with the results displayed in Fig. 5.2.

VIII.6 Implementation of Implicit Methods

Symplectic methods for general Hamiltonian equations are implicit, and so are symmetric methods for general reversible systems. Also, when we consider variable step size extensions as described in Sections VIII.3 and VIII.2, we are led to nonlinear equations. The efficient numerical solution of such nonlinear equations is the main difficulty in an implementation of implicit methods. Notice that in the context of geometric integration there is no need of ad-hoc strategies for step size and order selection, so that the remaining parts of a computer code are more or less straightforward.

In the following we discuss the numerical solution of the nonlinear system defined by an implicit Runge–Kutta method. We have the Gauss methods of Sect. II.1.3 in mind which are symplectic and symmetric. An extension of the ideas to partitioned Runge–Kutta methods and to Nyström methods is obvious. For simplicity of notation we consider autonomous differential equations $\dot y = f(y)$, and we write the nonlinear system of Definition II.1.1 in the form

$$Z_{in} - h \sum_{j=1}^{s} a_{ij} f\big(y_n + Z_{jn}\big) = 0, \qquad i = 1, \ldots, s. \tag{6.1}$$

The unknown variables are $Z_{1n}, \ldots, Z_{sn}$, and the equivalence of the two formulations is via the relation $k_i = f\big(y_n + Z_{in}\big)$. The numerical solution after one step can be expressed as

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f\big(y_n + Z_{in}\big). \tag{6.2}$$

For implicit Runge–Kutta methods the equations (6.1) represent a nonlinear system that has to be solved iteratively. We discuss the choice of good starting approximations for Z_{in} as well as different nonlinear equation solvers (fixed-point iteration, modified Newton methods).

VIII.6.1 Starting Approximations

The most simple approximations to the solution Z_{in} of (6.1) are $Z_{in}^0 = 0$ or $Z_{in}^0 = hc_i f(y_n)$ where $c_i = \sum_{j=1}^{s} a_{ij}$. They are, however, not very accurate and we will try to exploit the information of previous steps for improving them. There are essentially two possibilities: either use only the information of the last step $y_{n-1} \mapsto y_n$ (methods (A) and (B) below), or consider a fixed i and use the interpolation polynomial that passes through $Z_{i,n-l}$ for $l = 1, 2, \ldots$ (method (C)). Let us separately discuss these two approaches.

(A) Use of Continuous Output. Consider the polynomial $w_{n-1}(t)$ of degree s that interpolates the values (t_{n-1}, y_{n-1}) and $(t_{n-1}+c_i h, Y_{i,n-1})$ for $i = 1, \ldots, s$, where $Y_{i,n-1} = y_{n-1}+Z_{i,n-1}$ is the argument in (6.1) of the previous step. For collocation methods (such as Gauss methods) $w_{n-1}(t)$ is the collocation polynomial, and we know from Lemma II.1.6 that on compact intervals

$$w_{n-1}(t) - y(t) = \mathcal{O}(h^{q+1}) \tag{6.3}$$

with $q = s$, where $y(t)$ denotes the solution of $\dot y = f(y)$ satisfying $y(t_{n-1}) = y_{n-1}$. For Runge–Kutta methods that are not collocation methods, (6.3) holds with q defined by the condition $C(q)$ of (II.1.11). Since the solution of $\dot y = f(y)$ passing through $y(t_n) = y_n$ is $\mathcal{O}(h^{p+1})$ close to $y(t)$ with $p \ge q$, we have $w_n(t) = w_{n-1}(t) + \mathcal{O}(h^{q+1})$ and the computable value

$$Z_{in}^0 = Y_{in}^0 - y_n, \qquad Y_{in}^0 = w_{n-1}(t_n + c_i h) \tag{6.4}$$

serves as starting approximation for (6.1) with an error of size $\mathcal{O}(h^{q+1})$. This approach is standard in variable step size implementations of implicit Runge–Kutta methods (cf. Sect. IV.8 of Hairer & Wanner (1996)). Since $w_{n-1}(t) - y_{n-1}$ is a linear combination of the $Z_{i,n-1} = Y_{i,n-1} - y_{n-1}$, it follows from (6.1) that it is also a linear combination of $hf(Y_{i,n-1})$, so that

$$Y_{in}^0 = y_{n-1} + h \sum_{j=1}^{s} \beta_{ij} f(Y_{j,n-1}). \tag{6.5}$$

For a constant step size implementation, the β_{ij} depend only on the method coefficients and can be computed in advance as the solution of the linear Vandermonde type system

$$\sum_{j=1}^{s} \beta_{ij} c_j^{k-1} = \frac{(1+c_i)^k}{k}, \qquad k = 1, \ldots, s \tag{6.6}$$

(see Exercise 2). For collocation methods and for methods with $q \geq s-1$ the coefficients β_{ij} from (6.6) are optimal in the sense that they are the only ones making (6.5) an sth order approximation to the solution of (6.1). For $q < s-1$, more complicated order conditions have to be considered (Sand 1992).

(B) Starting Algorithms Using Additional Function Evaluations. In particular for high order methods where s is relatively large, a much more accurate starting approximation can be constructed with the aid of a few additional function evaluations. Such starting algorithms have been investigated by Laburta (1997), who presents coefficients for the Gauss methods up to order 8 in Laburta (1998).

The idea is to use starting approximations of the form

$$Y_{in}^0 = y_{n-1} + h \sum_{j=1}^{s} \beta_{ij} f(Y_{j,n-1}) + h \sum_{j=1}^{m} \nu_{ij} f(Y_{s+j,n-1}), \tag{6.7}$$

where $Y_{1,n-1}, \ldots, Y_{s,n-1}$ are the internal stages of the basic implicit Runge–Kutta method (with coefficients c_i, a_{ij}, b_j), and the additional internal stages are computed from

$$Y_{s+i,n-1} = y_{n-1} + h \sum_{j=1}^{s+i-1} \mu_{ij} f(Y_{j,n-1}).$$

For a fixed i, we interpret Y_{in}^0 as the result of the explicit Runge–Kutta method with coefficients of the right tableau of

$$\text{exact } i\text{th stage} \quad \begin{array}{c|cc} c & A & \\ \mathbb{1}+c & B & A \\ \hline & b^T & a_i^T \end{array} \qquad\qquad \text{approximate} \quad \begin{array}{c|cc} c & A & \\ \mu & M_1 & M_2 \\ \hline & \beta_i^T & \nu_i^T \end{array} \tag{6.8}$$

Here, $(M_1, M_2) = M = (\mu_{jk})$, $\mu_j = \sum_{k=1}^{s+j-1} \mu_{jk}$, and c, μ, β_i, ν_i are the vectors composed of $c_j, \mu_j, \beta_{ij}, \nu_{ij}$, respectively. The exact stage values Y_{in} are interpreted as the result of the Runge–Kutta method with coefficients given in the left tableau of (6.8). The entries of the vectors $\mathbb{1}, b$ and a_i are $1, b_j$ and a_{ij}, respectively, and B is the matrix whose rows are all equal to b^T.

If the order conditions (see Sect. III.1) for the two Runge–Kutta methods of (6.8) give the same result for all trees with $\leq r$ vertices, we get an approximation of order r, i.e., $Y_{in}^0 - Y_{in} = \mathcal{O}(h^{r+1})$. For the bushy tree $\tau_k = [\bullet, \ldots, \bullet]$ with k vertices we have

$$\sum_{j=1}^{s} \beta_{ij} c_j^{k-1} + \sum_{j=1}^{m} \nu_{ij} \mu_j^{k-1} = \sum_{j=1}^{s} b_j c_j^{k-1} + \sum_{j=1}^{s} a_{ij} (1+c_j)^{k-1}. \tag{6.9}$$

Notice that for collocation methods (such as the Gauss methods) the condition $C(s)$ reduces the right-hand expression of this equation to $(1+c_i)^k/k$ for $k \le s$. For $m = 0$, these conditions are thus equivalent to (6.6).

For the tree $[\tau_k] = \big[[\bullet, \ldots, \bullet]\big]$ with $k+1$ vertices we get the condition

$$\begin{aligned}\sum_{j,l=1}^{s} \beta_{ij} a_{jl} c_l^{k-1} + \sum_{j=1}^{m} \nu_{ij} \Big(\sum_{l=1}^{s} \mu_{jl} c_l^{k-1} + \sum_{l=1}^{m} \mu_{j,s+l} \mu_l^{k-1} \Big) \\ = \sum_{j,l=1}^{s} b_j a_{jl} c_l^{k-1} + \sum_{j,l=1}^{s} a_{ij} \Big(b_l c_l^{k-1} + a_{jl} (1+c_l)^{k-1} \Big).\end{aligned} \tag{6.10}$$

We now assume that the Runge–Kutta method corresponding to the right tableau of (6.8) satisfies condition $C(s)$. This means that the method (c, A, b) is a collocation method, and that the coefficients μ_{ij} have to be computed from the linear system

$$\sum_{j=1}^{s+i-1} \mu_{ij} c_j^{k-1} = \frac{\mu_i^k}{k}, \qquad k = 1, \ldots, s. \tag{6.11}$$

The method corresponding to the left tableau of (6.8) then also satisfies $C(s)$. Consequently, the order conditions are simplified considerably, and it follows from Sect. III.1 that Y_{in}^0 is an approximation to the exact stage value Y_{in} of order $s+1$ or $s+2$ if the following conditions hold:

$$\begin{aligned}&\text{order } s+1 \qquad \text{if (6.9) for } k = 1, \ldots, s+1; \\ &\text{order } s+2 \qquad \text{if (6.9) for } k = 1, \ldots, s+2, \text{ and (6.10) for } k = s+1.\end{aligned} \tag{6.12}$$

For an approximation of *order* $s+1$ we put $m = 1$, we arbitrarily choose μ_1, we compute μ_{1j} from (6.11), and the coefficients β_{ij} and ν_{i1} from (6.9) with $k = 1, \ldots, s+1$. A reasonable choice for the free parameter is $\mu_1 \in [1, 2]$ (in our computations we take $\mu_1 = 1.75$ for $s = 2, 4$, and $\mu_1 = 1.8$ for $s = 6$.[1]

For an approximation of *order* $s+2$ we put $m = 3$. One of the three additional function evaluations can be saved if we put $\mu_1 = 0$ and $\mu_2 = 1$. This implies $Y_{s+1,n-1} = y_{n-1}$ and $Y_{s+2,n-1} = y_n$, so that the evaluation of $f(Y_{s+1,n-1})$ is already available from computations for the preceding step (FSAL technique, "first same as last"). In our experiments we take $\mu_3 = 1.6$ for $s = 2$, $\mu_3 = 1.65$ for $s = 4$, and $\mu_3 = 1.75$ for $s = 6$. The coefficients $\mu_{ij}, \beta_{ij}, \nu_{ij}$ are then obtained as the solution of Vandermonde like linear systems.

For an implementation it is more convenient to work with the quantities $Z_{in}^0 = Y_{in}^0 - y_n$ and to write (6.7) in the form

[1] Laburta (1997) proposes to consider $m = 2$, $\mu_1 = 0$, $\mu_2 = 1$ (apart from the first step this also needs only one additional function evaluation per step), and to optimize free parameters by satisfying the order conditions for some trees with one order higher.

$$Z_{in}^0 = h\sum_{j=1}^{s}\alpha_{ij}f(Y_{j,n-1}) + h\sum_{j=1}^{m}\nu_{ij}f(Y_{s+j,n-1}) \tag{6.13}$$

with $\alpha_{ij} = \beta_{ij} - b_j$.

(C) Equistage Approximation. From the implicit function theorem, applied to the nonlinear system (6.1), we know that $Z_{in} = z(y_n, h)$, where the function $z(y,h)$ is as smooth as $f(y)$. Furthermore, since on compact intervals the global error of a one-step method permits an asymptotic expansion in powers of h, we have $y_{n-l} = y_N(t_{n-l}, h) + \mathcal{O}(h^{N+1})$ with $y_N(t,h) = y(t) + h^p e_p(t) + \ldots + h^N e_N(t)$ (the value of N can be chosen arbitrarily large if $f(y)$ is sufficiently smooth). Consequently, $Z_{i,n-l}$ is $\mathcal{O}(h^{N+1})$ close to the smooth function $z\big(y_N(t,h),h\big)$ at $t = t_n - lh$. Let $\zeta_i(t)$ be the polynomial of degree $k-1$ defined by $\zeta_i(t_{n-l}) = Z_{i,n-l}$ for $l = 1,\ldots,k$. Then, the value

$$Z_{in}^0 = \zeta_i(t_n) \tag{6.14}$$

yields a $\mathcal{O}(h^{k+1})$ approximation to the solution of (6.1). This interpolation procedure was first proposed by In't Hout (1992) for the numerical solution of delay differential equations. For the iterative solution of the nonlinear Runge–Kutta equations (6.1), the starting approximation (6.14) is proposed and analyzed by Calvo (2002).

The implementation of this approach is very simple. Using Newton's interpolation formula we have

$$Z_{in}^0 \;=\; Z_{i,n-1} + \nabla Z_{i,n-1} + \ldots + \nabla^{k-1} Z_{i,n-1} \tag{6.15}$$

with backward differences given by $\nabla Z_{i,n} = Z_{i,n} - Z_{i,n-1}$, $\nabla^2 Z_{i,n} = \nabla Z_{i,n} - \nabla Z_{i,n-1}$, etc.

Numerical Study of Starting Approximations. We consider the Kepler problem with eccentricity $e = 0.6$ and initial values such that the period is 2π. With many different step sizes $h = 2\pi/N$ we compute $N+1$ steps with the Gauss method of order $p = 2s$ ($p = 4, 8, 12$). In the last step we compute the different starting approximations and their error $(\sum_{i=1}^{s}\|Z_{in} - Z_{in}^0\|^2)^{1/2}$ as a function of the step size h. The result is plotted in Fig. 6.1. There, the pictures also contain the global errors after one period. They allow us to localize the values of h, which are of practical interest.

We observe that the equistage approximation (6.15) also behaves like $\mathcal{O}(h^{k+1})$ when $k+1$ is larger than the order of the integrator. However, due to the increasing error constants, the accuracy is improved only for small step sizes. An optimal k could be estimated by checking the decrease of the backward differences $\|\nabla^j Z_{i,n-1}\|$. The error of the starting approximation obtained from the continuous output behaves like $\mathcal{O}(h^{s+1})$ (for the Gauss methods) and, in contrast to the equistage approximation, improves with increasing order. The approximations (6.7) of order $s+1$ and $s+2$ are a clear improvement. As a conclusion we find that for this example the equistage approximation (which is free from additional function evaluations) is preferable only for $s = 2$ (order 4). For higher order, the approximation

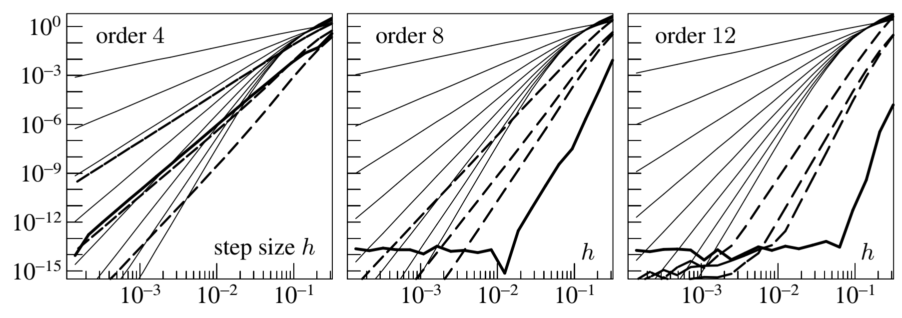


Fig. 6.1. Errors of starting approximations for Gauss methods as functions of the step size h: thick dashed lines for the extrapolated continuous output (6.4) and for the approximations (6.7) of order $s + 1$ and $s + 2$; thin solid lines for the equistage approximation (6.15) with $k = 0, 1, \ldots, 7$; the thick solid line represents the global error of the method after one period

obtained from (6.7) is significantly more accurate and so it is worthwhile to spend these two additional function evaluations per step.

VIII.6.2 Fixed-Point Versus Newton Iteration

Finally we investigate the iterative solution of the nonlinear Runge–Kutta system (6.1). We discuss fixed-point and Newton-like iterations, and we compare their efficiency to the use of composition methods.

Fixed-Point Iteration. This is the most simple and most natural iteration for the solution of (6.1). With any starting approximation Z_{in}^0 from Sect. VIII.6.1 it reads

$$Z_{in}^{k+1} = h \sum_{j=1}^{s} a_{ij} f\big(y_n + Z_{jn}^k\big), \qquad i = 1, \ldots, s. \tag{6.16}$$

In the case where the entries of the Jacobian matrix $f'(y)$ are not excessively large (nonstiff problems) and that the step size is sufficiently small, this iteration converges for $k \to \infty$ to the solution of (6.1). Usually, the iteration is stopped if a certain norm of the differences $Z_{in}^{k+1} - Z_{in}^k$ is sufficiently small. We then use Z_{in}^k in the update formula (6.2) so that no additional function evaluation is required.

For a numerical study of the convergence of this iteration, we consider the Kepler problem with eccentricity $e = 0.6$ and initial values as in the preceding experiments (period of the solution is 2π). We apply the Gauss methods of order 4, 8, and 12 with various step sizes. For the integration over one period we show in Table 6.1 the total number of function evaluations, the mean number of required iterations per step, and the global error at the endpoint of integration. As a stopping criterion for the fixed-point iteration we check whether the norm of the difference of two successive approximations is smaller than 10^{-16} (roundoff unit in double precision). As a starting approximation Z_{in}^0 we use (6.15) with $k = 8$ for the method of order 4,

Table 6.1. Statistics of Gauss methods (total number of function evaluations, number of fixed-point iterations per step, and the global error at the endpoint) for computations of the Kepler problem over one period with $e = 0.6$

Fixed-point iteration (general problems)

Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	803 16.1 $9.2 \cdot 10^{-2}$	1 043 10.4 $1.7 \cdot 10^{-2}$	1 393 7.0 $1.3 \cdot 10^{-3}$	1 825 4.6 $8.4 \cdot 10^{-5}$	2 319 2.9 $5.3 \cdot 10^{-6}$
order 8	1 021 9.7 $1.1 \cdot 10^{-3}$	1 455 6.8 $6.9 \cdot 10^{-7}$	2 091 4.7 $3.6 \cdot 10^{-9}$	3 007 3.3 $1.8 \cdot 10^{-11}$	4 183 2.1 $6.9 \cdot 10^{-14}$
order 12	1 297 8.3 $2.7 \cdot 10^{-6}$	1 731 5.4 $8.0 \cdot 10^{-11}$	2 311 3.5 $2.7 \cdot 10^{-14}$	3 441 2.5 $\leq$ roundoff	5 917 2.1 $\leq$ roundoff

and the approximation (6.7) of order $s + 2$ for the methods of orders 8 and 12. The coefficients are those presented after equation (6.12).

Since the starting approximations are more accurate for small h, the number of necessary iterations decreases drastically. In particular, for the 4th order method we need about 16 iterations per step for $h = 2\pi/25$, but at most 2 iterations when $h \leq 2\pi/800$. If one is interested in high accuracy computations (e.g., long-time simulations in astronomy), for which the error over one period is not larger than 10^{-10}, Table 6.1 illustrates that high order methods ($p \geq 12$) are most efficient.

Newton-Type Iterations. A standard technique for solving nonlinear equations is Newton's method or some modification of it. Writing the nonlinear system (6.1) of an implicit Runge–Kutta method as $F(Z) = 0$ with $Z = (Z_{1n}, \dots, Z_{sn})^T$, the Newton iteration is

$$Z^{k+1} = Z^k - M^{-1} F(Z^k), \tag{6.17}$$

where M is some approximation to the Jacobian matrix $F'(Z^k)$. Since the solution Z of the nonlinear system is $\mathcal{O}(h)$ close to zero, it is common to use $M = F'(0)$ so that the matrix M is independent of the iteration index k. In our special situation we get

$$M = I \otimes I - hA \otimes J \tag{6.18}$$

with $J = f'(y_n)$. Here, I denotes the identity matrix of suitable dimension, and A is the Runge–Kutta matrix.

We repeat the experiment of Table 6.1 with modified Newton iterations instead of fixed-point iterations. The result is shown in Table 6.2. We have suppressed the error at the end of the period, because it is the same as in Table 6.1. As expected, the convergence is faster (i.e., the number of iterations per step is smaller) so that the total number of function evaluations is reduced. However, we do not see in this table that we computed at every step the Jacobian $f'(y_n)$ and an LR-decomposition of the matrix M. Even if we exploit the tensor product structure in (6.18) as explained

Table 6.2. Statistics of Gauss methods (total number of function evaluations, number of iterations per step) for computations of the Kepler problem over one period with $e = 0.6$

Modified Newton iteration (general problems)

Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	383 7.7	511 5.1	765 3.8	1 125 2.8	1 677 2.1
order 8	597 5.5	883 3.9	1 387 3.0	2 307 2.4	3 667 1.8
order 12	763 4.7	1 095 3.3	1 717 2.5	3 003 2.2	5 689 2.0

in Hairer & Wanner (1996, Sect. IV.8), the cpu time is now considerably larger. Further improvements are possible, if the Jacobian of f and hence also the LR-decomposition of M is frozen over a couple of steps. But all these efforts can hardly beat (in cpu time) the straightforward fixed-point iterations. In accordance with the experience of Sanz-Serna & Calvo (1994, Sect. 5.5) we recommend in general the use of fixed-point iterations.

Separable Systems and Second Order Differential Equations. Many interesting differential equations are of the form

$$\dot{\eta} = f(y), \qquad \dot{y} = g(\eta). \tag{6.19}$$

For example, the second order differential equation $\ddot{y} = f(y)$ is obtained by putting $g(\eta) = \eta$. Also Hamiltonian systems with separable Hamiltonian $H(p,q) = T(p) + U(q)$ are of the form (6.19).

For this particular system the Runge–Kutta equations (6.1) become

$$\zeta_{in} - h\sum_{j=1}^{s} a_{ij} f(y_n + Z_{jn}) = 0, \qquad Z_{in} - h\sum_{j=1}^{s} a_{ij} g(\eta_n + \zeta_{jn}) = 0.$$

In this case we can still do better: instead of the standard fixed-point iteration (6.16) we apply a Gauss-Seidel like iteration

$$\zeta_{in}^{k+1} = h\sum_{j=1}^{s} a_{ij} f(y_n + Z_{jn}^{k}), \qquad Z_{in}^{k+1} = h\sum_{j=1}^{s} a_{ij} g(\eta_n + \zeta_{jn}^{k+1}), \tag{6.20}$$

which is explicit for separable systems (6.19). Notice that the starting approximations have to be computed only for ζ_{in}. Those for Z_{in} are then obtained by (6.20) with $k + 1 = 0$.

For second order differential equations $\ddot{y} = f(y)$, where $g(\eta) = \eta$, this iteration becomes

$$Z_{in}^{k+1} = hc_i\eta_n + h^2\sum_{j=1}^{s} \widehat{a}_{ij} f(y_n + Z_{jn}^{k}), \tag{6.21}$$

Table 6.3. Statistics of iterations (6.20) for Gauss methods (total number of function evaluations, number of iterations per step) for computations of the Kepler problem over one period with $e = 0.6$

Fixed-point iteration (separable problems)

Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	437 8.7	603 6.0	857 4.3	1 201 3.0	1 717 2.1
order 8	613 5.6	923 4.1	1 427 3.1	2 339 2.4	3 647 1.8
order 12	781 4.9	1 131 3.4	1 741 2.6	3 027 2.2	5 677 2.0

where $c_i = \sum_{j=1}^{s} a_{ij}$ and $\widehat{a}_{ij}$ are the entries of the square A^2 of the Runge–Kutta matrix (any Nyström method could be applied as well). Due to the factor h^2 in (6.21) we expect this iteration to converge about twice as fast as the standard fixed-point iteration.

The Kepler problem is a second order differential equation, so that the iteration (6.21) can be applied. In analogy to the previous tables we present in Table 6.3 the statistics of such an implementation of the Gauss methods. We observe that for relatively large step sizes the number of iterations required per step in nearly halved (compared to Table 6.1). For high accuracy requirements the number of necessary iterations is surprisingly small, and the question arises whether such an implementation can compete with high order explicit composition methods.

Comparison Between Implicit Runge–Kutta and Composition Methods. We consider second order differential equations $\ddot{y} = f(y)$, so that composition methods based on the explicit Störmer–Verlet scheme can be applied. We use the coefficients of method (V.3.14) which has turned out to be excellent in the experiments of Sect. V.3.2. It is a method of order 8 and uses 17 function evaluations per integration step.

We compare it with the Gauss methods of order 8 and 12 (i.e., $s = 4$ and $s = 6$). As a starting approximation for the solution of the nonlinear system (6.1) we use (6.7) with $m = 3$, $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 1.75$, μ_{ij} chosen such that (6.11) holds for $k = 1, \ldots, s + i - 1$, and β_{ij}, ν_{ij} such that order $s + 2$ is obtained. Since we are concerned with second order differential equations, we apply the iterations (6.20) until the norm of the difference of two successive approximations is below 10^{-17}.

For both classes of methods we use compensated summation (Algorithm 5.1), which permits us to reduce rounding errors. For composition methods we apply this technique for all updates of the basic integrator. For Runge–Kutta methods, we use it for adding the increment to y_n and also for computing the sum $\sum_{i=1}^{s} b_i k_i$.

The work–precision diagrams of the comparison are given in Fig. 6.2. The upper pictures correspond to the Kepler problem with $e = 0.6$ and an integration over 100 periods; the lower pictures correspond to the outer solar system with data given in Sect. I.2.4 and an integration over $500\,000$ earth days. The left pictures show the

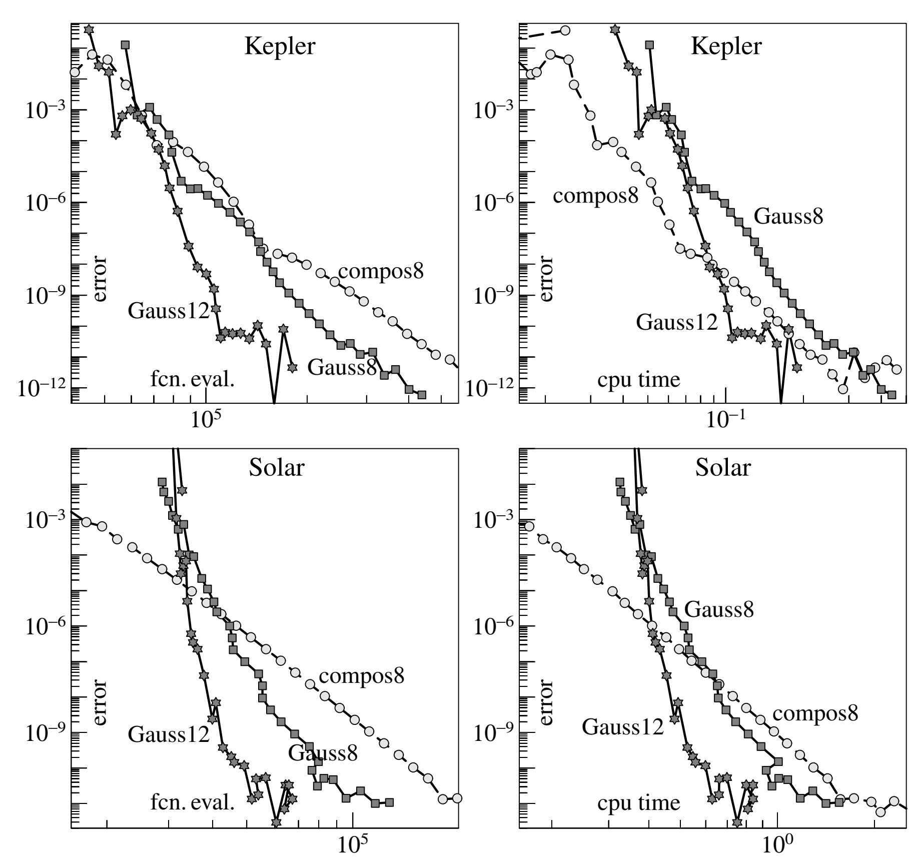


Fig. 6.2. Work–precision diagrams for two problems (Kepler and outer solar system) and three numerical integrators (composition method with coefficients of method (V.3.14) based on the explicit Störmer–Verlet scheme and the Gauss methods of orders 8 and 12)

Euclidean norm of the error at the end of the integration interval as a function of total numbers of function evaluations required for the integration; the pictures to the right present the same error as a function of the cpu times (with optimizing compiler on a SunBlade 100 workstation). We can draw the following conclusions from this experiment:

- the implementation of composition methods based on the Störmer–Verlet scheme is extremely easy; that of implicit Runge–Kutta methods is slightly more involved because it requires a stopping criterion for the fixed-point iterations;
- the overhead (total cpu time minus that used for the function evaluations) is much higher for the implicit Runge–Kutta methods; this is seen from the fact that implicit Runge–Kutta methods require less function evaluations for a given accuracy, but often more cpu time;
- among the two Gauss methods, the higher order method is more efficient for all precisions of practical interest;

- for very accurate computations (say, in quadruple precision), high order Runge–Kutta methods are more efficient than composition methods;
- much of the computation in the Runge–Kutta code can be done in parallel (e.g., the s function evaluations of a fixed-point iteration); composition methods do not have this potential;
- implicit Runge–Kutta methods can be applied to general (non-separable) differential equations, and the cost of the implementation is at most twice as large; if one is obliged to use an implicit method as the basic method for composition, many advantages of composition methods are lost.

Both classes of methods (composition and implicit Runge–Kutta) are of interest in the geometric integration of differential equations. Each one has its advantages and disadvantages.

Fortran codes of these computations are available on the Internet under the homepage <http://www.unige.ch/math/folks/hairer>. A Matlab version of these codes is described in E. & M. Hairer (2003).

VIII.7 Exercises

1. Consider a one-step method applied to a Hamiltonian system. Give a probabilistic proof of the property that the error of the numerical Hamiltonian due to roundoff grows like $\mathcal{O}(\sqrt{t}\, eps)$.
2. Prove that the collocation polynomial can be written as

$$w_n(t) = y_n + h \sum_{i=1}^{s} \beta_i(t)\, f(Y_{in}),$$

   where the polynomials $\beta_i(t)$ are a solution of

$$\sum_{j=1}^{s} \beta_j(t)\, c_j^{k-1} \;=\; \frac{t^k}{k}.$$

3. Apply your favourite code to the Kepler problem and to the outer solar system with data as in Fig. 6.2. Plot a work-precision diagram.
   Remark. Figure 7.1 shows our results obtained with the 8th order Runge–Kutta code Dop853 (Hairer, Nørsett & Wanner 1993) compared to an 8th order composition method. Rounding errors are more pronounced for Dop853, because compensated summation is not applied. Computations on shorter time intervals and comparisons of required function evaluations would be more in favour for Dop853. It is also of interest to consider high order Runge–Kutta Nyström methods.
4. Consider starting approximations

$$Y_{in}^0 = y_{n-2} + h \sum_{j=1}^{s} \beta_{ij}^{(2)} f(Y_{j,n-2}) + h \sum_{j=1}^{s} \beta_{ij}^{(1)} f(Y_{j,n-1}) \tag{7.1}$$

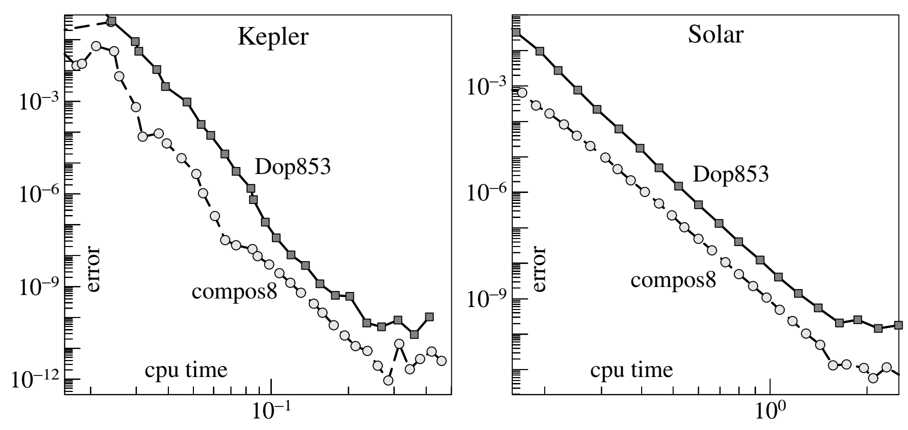


Fig. 7.1. Work–precision diagrams for the explicit, variable step size Runge–Kutta code Dop853 applied to two problems (Kepler and outer solar system). For a comparison, the results of Fig. 6.2 for the composition method are included

which use the internal stages of two consecutive steps without any additional function evaluation. What are the conditions such that (7.1) is of order $s+1$, of order $s+2$?
Compare the efficiency of these formulas with the algorithms (A) and (B) of Sect. VIII.6.1.

5. Prove that for a second order differential equation $\ddot{y} = f(y)$ (more precisely, for $\dot{y} = z, \dot{z} = f(y)$) the application of the s-stage Gauss method gives

$$\begin{aligned} y_{n+1} &= y_n + h\dot{y}_n + h^2 \sum_{i=1}^{s} b_i(1-c_i) f(y_n + Z_{in}) \\ \dot{y}_{n+1} &= \dot{y}_n + h \sum_{i=1}^{s} b_i f(y_n + Z_{in}), \end{aligned}$$

where Z_{in} is obtained from the iteration (6.21).
Hint. The coefficients of the Gauss methods satisfy $\sum_j b_j a_{ji} = b_i(1-c_i)$ for all i.

Chapter IX. Backward Error Analysis and Structure Preservation

> One of the greatest virtues of backward analysis … is that when it *is* the appropriate form of analysis it tends to be very markedly superior to forward analysis. Invariably in such cases it has remarkable formal simplicity and gives deep insight into the stability (or lack of it) of the algorithm. (J.H. Wilkinson, IMA Bulletin 1986)

The origin of backward error analysis dates back to the work of Wilkinson (1960) in numerical linear algebra. For the study of integration methods for ordinary differential equations, its importance was seen much later. The present chapter is devoted to this theory. It is very useful, when the qualitative behaviour of numerical methods is of interest, and when statements over very long time intervals are needed. The formal analysis (construction of the modified equation, study of its properties) gives already a lot of insight into numerical methods. For a rigorous treatment, the modified equation, which is a formal series in powers of the step size, has to be truncated. The error, induced by such a truncation, can be made exponentially small, and the results remain valid on exponentially long time intervals.

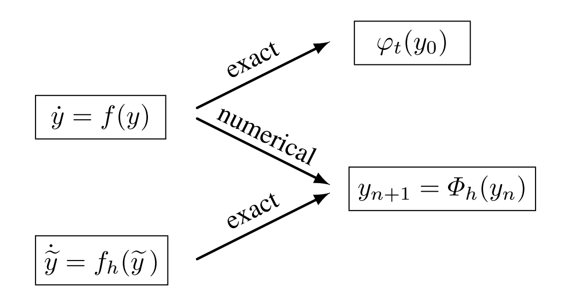
IX.1 Modified Differential Equation – Examples

Consider an ordinary differential equation

$$\dot{y} = f(y),$$

and a numerical method $\Phi_h(y)$ which produces the approximations

$$y_0, y_1, y_2, \ldots .$$



A forward error analysis consists of the study of the errors $y_1 - \varphi_h(y_0)$ (local error) and $y_n - \varphi_{nh}(y_0)$ (global error) in the solution space. The idea of backward error analysis is to search for a *modified differential equation* $\dot{\widetilde{y}} = f_h(\widetilde{y})$ of the form

$$\dot{\widetilde{y}} = f(\widetilde{y}) + hf_2(\widetilde{y}) + h^2 f_3(\widetilde{y}) + \ldots , \tag{1.1}$$

such that $y_n = \widetilde{y}(nh)$, and in studying the difference of the vector fields $f(y)$ and $f_h(y)$. This then gives much insight into the qualitative behaviour of the numerical solution and into the global error $y_n - y(nh) = \widetilde{y}(nh) - y(nh)$. We remark that the series in (1.1) usually diverges and that one has to truncate it suitably. The effect of such a truncation will be studied in Sect. IX.7. For the moment we content ourselves with a formal analysis without taking care of convergence issues. The idea of interpreting the numerical solution as the exact solution of a modified equation is common to many numerical analysts ("... This is possible since the map is the solution of some physical Hamiltonian problem which, in some sense, is close to the original problem", Ruth (1983), or "... the symplectic integrator creates a numerical Hamiltonian system that is close to the original ...", Gladman, Duncan & Candy 1991). A systematic study started with the work of Griffiths & Sanz-Serna (1986), Feng (1991), Sanz-Serna (1992), Yoshida (1993), Eirola (1993), Fiedler & Scheurle (1996), and many others.

For the computation of the modified equation (1.1) we put $y := \widetilde{y}(t)$ for a fixed t, and we expand the solution of (1.1) into a Taylor series

$$\begin{aligned} \widetilde{y}(t+h) &= y + h\big(f(y) + hf_2(y) + h^2 f_3(y) + \ldots\big) \\ &+ \frac{h^2}{2!}\big(f'(y) + hf_2'(y) + \ldots\big)\big(f(y) + hf_2(y) + \ldots\big) + \ldots . \end{aligned} \tag{1.2}$$

We assume that the numerical method $\Phi_h(y)$ can be expanded as

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \ldots \tag{1.3}$$

(the coefficient of h is $f(y)$ for consistent methods). The functions $d_j(y)$ are known and are typically composed of $f(y)$ and its derivatives. For the explicit Euler method we simply have $d_j(y) = 0$ for all $j \geq 2$. In order to get $\widetilde{y}(nh) = y_n$ for all n, we must have $\widetilde{y}(t+h) = \Phi_h(y)$. Comparing like powers of h in the expressions (1.2) and (1.3) yields recurrence relations for the functions $f_j(y)$, namely,

$$\begin{aligned} f_2(y) &= d_2(y) - \frac{1}{2!} f'f(y) \\ f_3(y) &= d_3(y) - \frac{1}{3!}\Big(f''(f,f)(y) + f'f'f(y)\Big) - \frac{1}{2!}\Big(f'f_2(y) + f_2'f(y)\Big). \end{aligned} \tag{1.4}$$

Example 1.1. Consider the scalar differential equation

$$\dot{y} = y^2, \qquad y(0) = 1 \tag{1.5}$$

with exact solution $y(t) = 1/(1-t)$. It has a singularity at $t = 1$. We apply the explicit Euler method $y_{n+1} = y_n + hf(y_n)$ with step size $h = 0.02$. The picture in Fig. 1.1 presents the exact solution (dashed curve) together with the numerical solution (bullets). The above procedure for the computation of the modified equation, implemented as a Maple program (see Hairer & Lubich 2000) gives

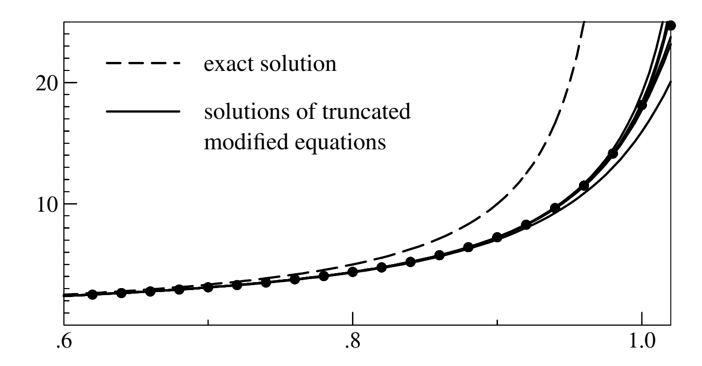


Fig. 1.1. Solutions of the modified equation for the problem (1.5)

```
> fcn := y -> y^2:
> nn := 6:
> fcoe[1] := fcn(y):
> for n from 2 by 1 to nn do
>    modeq := sum(h^j*fcoe[j+1], j=0..n-2):
>    diffy[0] := y:
>    for i from 1 by 1 to n do
>       diffy[i] := diff(diffy[i-1],y)*modeq:
>    od:
>    ytilde := sum(h^k*diffy[k]/k!, k=0..n):
>    res := ytilde-y-h*fcn(y):
>    tay := convert(series(res,h=0,n+1),polynom):
>    fcoe[n] := -coeff(tay,h,n):
> od:
> simplify(sum(h^j*fcoe[j+1], j=0..nn-1));
```

Its output is

$$\dot{\widetilde{y}} = \widetilde{y}^2 - h\widetilde{y}^3 + h^2\,\frac{3}{2}\,\widetilde{y}^4 - h^3\,\frac{8}{3}\,\widetilde{y}^5 + h^4\,\frac{31}{6}\,\widetilde{y}^6 - h^5\,\frac{157}{15}\,\widetilde{y}^7 \pm \ldots\,. \tag{1.6}$$

The above picture also presents the solution of the modified equation, when truncated after 1,2,3, and 4 terms. We observe an excellent agreement of the numerical solution with the exact solution of the modified equation.

A similar program for the implicit midpoint rule (I.1.7) computes the modified equation

$$\dot{\widetilde{y}} = \widetilde{y}^2 + h^2\,\frac{1}{4}\,\widetilde{y}^4 + h^4\,\frac{1}{8}\,\widetilde{y}^6 + h^6\,\frac{11}{192}\,\widetilde{y}^8 + h^8\,\frac{3}{128}\,\widetilde{y}^{10} \pm \ldots\,, \tag{1.7}$$

and for the classical Runge–Kutta method of order 4 (left tableau of (II.1.8))

$$\dot{\widetilde{y}} = \widetilde{y}^2 - h^4\,\frac{1}{24}\,\widetilde{y}^6 + h^6\,\frac{65}{576}\,\widetilde{y}^8 - h^7\,\frac{17}{96}\,\widetilde{y}^9 + h^8\,\frac{19}{144}\,\widetilde{y}^{10} \pm \ldots\,. \tag{1.8}$$

We observe that the perturbation terms in the modified equation are of size $\mathcal{O}(h^p)$, where p is the order of the method. This is true in general.

Theorem 1.2. *Suppose that the method $y_{n+1} = \Phi_h(y_n)$ is of order p, i.e.,*

$$\Phi_h(y) = \varphi_h(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2}),$$

where $\varphi_t(y)$ denotes the exact flow of $\dot y = f(y)$, and $h^{p+1}\delta_{p+1}(y)$ the leading term of the local truncation error. The modified equation then satisfies

$$\dot{\widetilde y} = f(\widetilde y\,) + h^p f_{p+1}(\widetilde y\,) + h^{p+1} f_{p+2}(\widetilde y\,) + \ldots, \qquad \widetilde y(0) = y_0 \tag{1.9}$$

with $f_{p+1}(y) = \delta_{p+1}(y)$.

Proof. The construction of the functions $f_j(y)$ (see the beginning of this section) shows that $f_j(y) = 0$ for $2 \le j \le p$ if and only if $\Phi_h(y) - \varphi_h(y) = \mathcal{O}(h^{p+1})$. □

A first application of the modified equation (1.1) is the existence of an *asymptotic expansion of the global error*. Indeed, by the nonlinear variation of constants formula, the difference between its solution $\widetilde y(t)$ and the solution $y(t)$ of $\dot y = f(y)$ satisfies

$$\widetilde y(t) - y(t) = h^p e_p(t) + h^{p+1} e_{p+1}(t) + \ldots\,. \tag{1.10}$$

Since $y_n = \widetilde y(nh) + \mathcal{O}(h^N)$ for the solution of a truncated modified equation, this proves the existence of an asymptotic expansion in powers of h for the global error $y_n - y(nh)$.

A large part of this chapter studies properties of the modified differential equation, and the question of the extent to which structures (such as conservation of invariants, Hamiltonian structure) in the problem $\dot y = f(y)$ can carry over to the modified equation.

Example 1.3. We next consider the Lotka–Volterra equations

$$\dot q = q(p-1), \qquad \dot p = p(2-q),$$

and we apply (a) the explicit Euler method, and (b) the symplectic Euler method, both with constant step size $h = 0.1$. The first terms of their modified equations are

$$\begin{aligned}
\text{(a)} \qquad \dot q &= q(p-1) - \frac{h}{2}\, q(p^2 - pq + 1) + \mathcal{O}(h^2),\\
\dot p &= -p(q-2) - \frac{h}{2}\, p(q^2 - pq - 3q + 4) + \mathcal{O}(h^2),\\
\text{(b)} \qquad \dot q &= q(p-1) - \frac{h}{2}\, q(p^2 + pq - 4p + 1) + \mathcal{O}(h^2),\\
\dot p &= -p(q-2) + \frac{h}{2}\, p(q^2 + pq - 5q + 4) + \mathcal{O}(h^2).
\end{aligned}$$

Figure 1.2 shows the numerical solutions for initial values indicated by a thick dot. In the pictures to the left they are embedded in the exact flow of the differential equation, whereas in those to the right they are embedded in the flow of the modified differential equation, truncated after the h^2 terms. As in the first example, we observe an excellent agreement of the numerical solution with the exact solution of

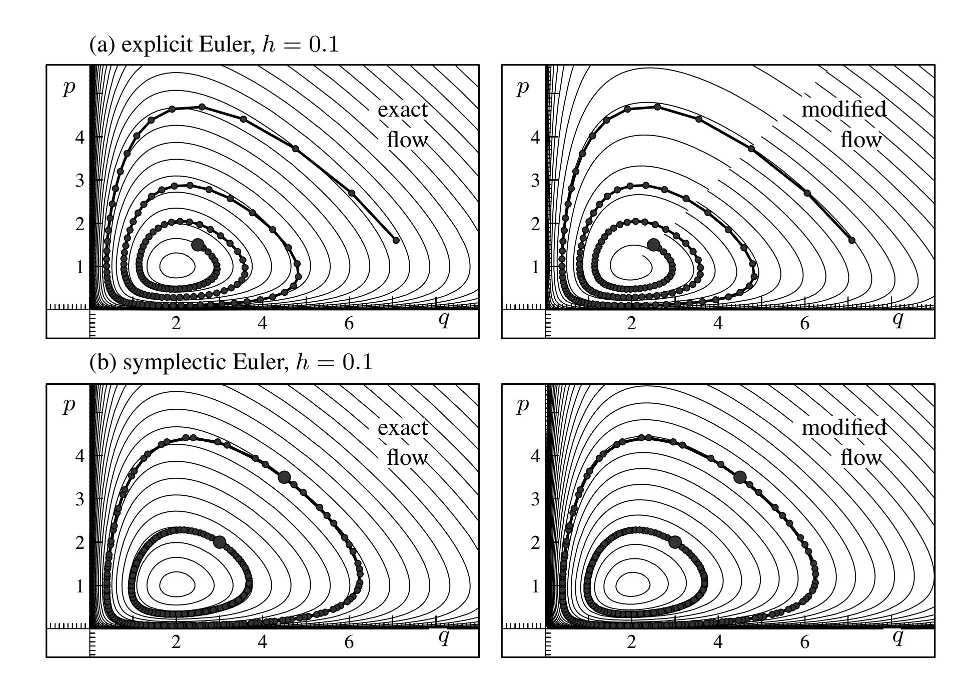


Fig. 1.2. Numerical solution compared to the exact and modified flows

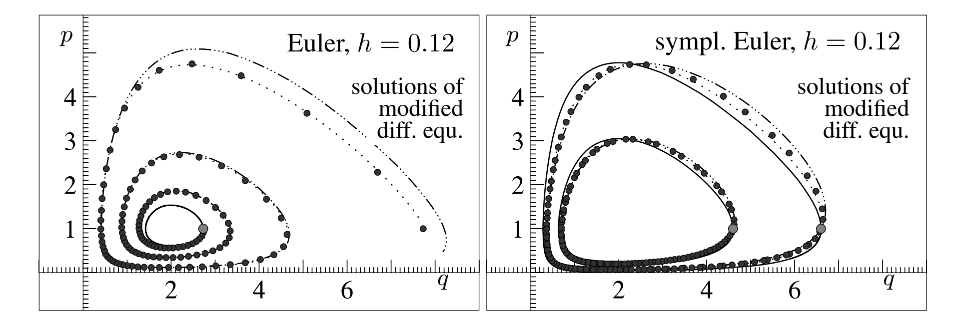


Fig. 1.3. Study of the truncation in the modified equation

the modified equation. For the symplectic Euler method, the solutions of the truncated modified equation are periodic, as is the case for the unperturbed problem (Exercise 5).

In Fig. 1.3 we present the numerical solution and the exact solution of the modified equation, once truncated after the h terms (dashed-dotted), and once truncated after the h^2 terms (dotted). The exact solution of the problem is included as a solid curve. This shows that taking more terms in the modified equation usually improves the agreement of its solution with the numerical approximation of the method.

Example 1.4. For a linear differential equation with constant coefficients

$$\dot{y} = Ay, \qquad y(0) = y_0$$

we consider numerical methods which yield $y_{n+1} = R(hA)y_n$, where $R(z)$ is the stability function (VI.4.9) of the method. In this case we get $y_n = R(hA)^n y_0$, so that $y_n = \widetilde{y}(nh)$, where $\widetilde{y}(t) = R(hA)^{t/h} y_0 = \exp\big(\frac{t}{h}\ln R(hA)\big)y_0$ is the solution of the modified differential equation

$$\dot{\widetilde{y}} = \frac{1}{h}\,\ln R(hA)\,\widetilde{y} = (A + hb_2A^2 + h^2b_3A^3 + \ldots)\,\widetilde{y} \tag{1.11}$$

with suitable constants $b_2, b_3, \ldots$. Since $R(z) = 1+z+\mathcal{O}(z^2)$ and $\ln(1+x) = x - x^2/2+\mathcal{O}(x^3)$ both have a positive radius of convergence, the series (1.11) converges for $|h| < h_0$ with some $h_0 > 0$. We shall see later that this is an exceptional situation. In general, the modified equation is a formal divergent series.

IX.2 Modified Equations of Symmetric Methods

In this and the following sections we investigate how the structure of the differential equation and geometric properties of the method are reflected in the modified differential equation. Here we begin by studying this question for symmetric/reversible methods.

Consider a numerical method Φ_h. Recall that its adjoint $y_{n+1} = \Phi_h^*(y_n)$ is defined by the relation $y_n = \Phi_{-h}(y_{n+1})$ (see Definition II.1.4).

Theorem 2.1 (Adjoint Methods). *Let $f_j(y)$ be the coefficient functions of the modified equation for the method Φ_h. Then, the coefficient functions $f_j^*(y)$ of the modified equation for the adjoint method Φ_h^* satisfy*

$$f_j^*(y) = (-1)^{j+1} f_j(y). \tag{2.1}$$

Proof. The solution $\widetilde{y}(t)$ of the modified equation for Φ_h^* has to satisfy $\widetilde{y}(t) = \Phi_{-h}\big(\widetilde{y}(t+h)\big)$ or, equivalently, $\widetilde{y}(t-h) = \Phi_{-h}(y)$ with $y := \widetilde{y}(t)$. We get (2.1) if we replace h with $-h$ in the formulas (1.1), (1.2) and (1.3). □

For symmetric methods we have $\Phi_h^* = \Phi_h$, implying $f_j^*(y) = f_j(y)$. We therefore get the following corollary to Theorem 2.1.

Theorem 2.2 (Symmetric Methods). *The coefficient functions of the modified equation of a symmetric method satisfy $f_j(y) = 0$ whenever j is even, so that (1.1) has an expansion in even powers of h.* □

This theorem explains the h^2-expansion in the modified equation (1.7) of the midpoint rule.

As a consequence of Theorem 2.2, the asymptotic expansion (1.10) of the global error is also in even powers of h. This property is responsible for the success of h^2-extrapolation methods.

Consider now a numerical method applied to a ρ-reversible differential equation as studied in Sect. V.1. Recall from Theorem V.1.5 that a symmetric method is ρ-reversible under the ρ-compatibility condition (V.1.4), which is satisfied for most numerical methods.

Theorem 2.3 (Reversible Methods). *Consider a ρ-reversible differential equation $\dot y = f(y)$ and a ρ-reversible numerical method $\Phi_h(y)$. Then, every truncation of the modified differential equation is again ρ-reversible.*

Proof. Let $f_j(y)$ be the jth coefficient of the modified equation (1.1) for Φ_h. The proof is by induction on j. So assume that for $j = 1, \ldots, r$, the vector field $f_j(y)$ is ρ-reversible, i.e.,

$$\rho \circ f_j = -f_j \circ \rho.$$

We show that the same relation holds also for $j = r+1$. By assumption, the truncated modified equation

$$\dot{\widetilde y} = f(\widetilde y\,) + hf_2(\widetilde y\,) + \ldots + h^{r-1} f_r(\widetilde y\,)$$

is ρ-reversible, so that by (V.1.2), it has a ρ-reversible flow $\varphi_{r,t}(y)$, that is, $\rho\circ\varphi_{r,t} = \varphi_{r,t}^{-1} \circ \rho$. By construction of the modified equation, we have

$$\Phi_h(y) = \varphi_{r,h}(y) + h^{r+1} f_{r+1}(y) + \mathcal{O}(h^{r+2}).$$

Since $\varphi_{r,h}(y) = y + \mathcal{O}(h)$, this implies

$$\Phi_h^{-1}(y) = \varphi_{r,h}^{-1}(y) - h^{r+1} f_{r+1}(y) + \mathcal{O}(h^{r+2}).$$

Since both Φ_h and $\varphi_{r,h}$ are ρ-reversible maps, these two relations yield $\rho \circ f_{r+1} = -f_{r+1} \circ \rho$ as desired. □

IX.3 Modified Equations of Symplectic Methods

We now present one of the most important results of this chapter. We consider a Hamiltonian system $\dot y = J^{-1}\nabla H(y)$ with an infinitely differentiable Hamiltonian $H(y)$, and we show that the modified equation of symplectic methods is also Hamiltonian.

IX.3.1 Existence of a Local Modified Hamiltonian

> ... if we neglect convergence questions then one can always find a formal integral ... (J. Moser 1968)

Theorem 3.1. *If a symplectic method $\Phi_h(y)$ is applied to a Hamiltonian system with a smooth Hamiltonian $H : \mathbb{R}^{2d} \to \mathbb{R}$, then the modified equation (1.1) is also Hamiltonian. More precisely, there exist smooth functions $H_j : \mathbb{R}^{2d} \to \mathbb{R}$ for $j = 2, 3, \ldots$, such that $f_j(y) = J^{-1}\nabla H_j(y)$.*

The following proof by induction, whose ideas can be traced back to Moser (1968), was given by Benettin & Giorgilli (1994) and Tang (1994). It can be extended to many other situations. We have already encountered its reversible version in the proof of Theorem 2.3.

Proof. Assume that $f_j(y) = J^{-1}\nabla H_j(y)$ for $j = 1, 2, \ldots, r$ (this is satisfied for $r = 1$, because $f_1(y) = f(y) = J^{-1}\nabla H(y)$). We have to prove the existence of a Hamiltonian $H_{r+1}(y)$. The idea is to consider the truncated modified equation

$$\dot{\widetilde{y}} = f(\widetilde{y}\,) + hf_2(\widetilde{y}\,) + \ldots + h^{r-1}f_r(\widetilde{y}\,), \tag{3.1}$$

which is a Hamiltonian system with Hamiltonian $H(y)+hH_2(y)+\ldots+h^{r-1}H_r(y)$. Its flow $\varphi_{r,t}(y_0)$, compared to that of (1.1), satisfies

$$\Phi_h(y_0) = \varphi_{r,h}(y_0) + h^{r+1}f_{r+1}(y_0) + \mathcal{O}(h^{r+2}),$$

and also

$$\Phi_h'(y_0) = \varphi_{r,h}'(y_0) + h^{r+1}f_{r+1}'(y_0) + \mathcal{O}(h^{r+2}).$$

By our assumption on the method and by the induction hypothesis, Φ_h and $\varphi_{r,h}$ are symplectic transformations. This, together with $\varphi_{r,h}'(y_0) = I + \mathcal{O}(h)$, therefore implies

$$J = \Phi_h'(y_0)^T J \Phi_h'(y_0) = J + h^{r+1}\Big(f_{r+1}'(y_0)^T J + J f_{r+1}'(y_0)\Big) + \mathcal{O}(h^{r+2}).$$

Consequently, the matrix $Jf_{r+1}'(y)$ is symmetric and the existence of $H_{r+1}(y)$ satisfying $f_{r+1}(y) = J^{-1}\nabla H_{r+1}(y)$ follows from the Integrability Lemma VI.2.7. This part of the proof is similar to that of Theorem VI.2.6. □

For Hamiltonians $H : D \to \mathbb{R}$ the statement of the above theorem remains valid with $H_j : D \to \mathbb{R}$ on domains $D \subset \mathbb{R}^{2d}$ on which the Integrability Lemma VI.2.7 is applicable. This is the case for simply connected domains D, but not in general (see the discussion after the proof of Lemma VI.2.7).

IX.3.2 Existence of a Global Modified Hamiltonian

By Lemma VI.5.3 every symplectic one-step method $\Phi_h : (p, q) \mapsto (P, Q)$ can be locally expressed in terms of a generating function $S(P, q, h)$ as

$$p = P + \frac{\partial S}{\partial q}(P, q, h), \qquad Q = q + \frac{\partial S}{\partial P}(P, q, h). \tag{3.2}$$

This property allows us to give an independent proof of Theorem 3.1 and in addition to show that the modified equation is Hamiltonian with $\widetilde{H}(p, q)$ defined on the same domain as the generating function. The following result is mentioned in Benettin & Giorgilli (1994) and in the thesis of Murua (1994), p. 100.

Theorem 3.2. *Assume that the symplectic method Φ_h has a generating function*

$$S(P,q,h) \;=\; h\,S_1(P,q) + h^2 S_2(P,q) + h^3 S_3(P,q) + \ldots \tag{3.3}$$

with smooth $S_j(P,q)$ defined on an open set D. Then, the modified differential equation is a Hamiltonian system with

$$\widetilde{H}(p,q) = H(p,q) + h\,H_2(p,q) + h^2 H_3(p,q) + \ldots\;, \tag{3.4}$$

where the functions $H_j(p,q)$ are defined and smooth on the whole of D.

Proof. By Theorem VI.5.7, the exact solution $(P,Q) = \big(\widetilde{p}(t),\widetilde{q}(t)\big)$ of the Hamiltonian system corresponding to $\widetilde{H}(p,q)$ is given by

$$p = P + \frac{\partial \widetilde{S}}{\partial q}(P,q,t), \qquad Q = q + \frac{\partial \widetilde{S}}{\partial P}(P,q,t),$$

where $\widetilde{S}$ is the solution of the Hamilton–Jacobi differential equation

$$\frac{\partial \widetilde{S}}{\partial t}(P,q,t) \;=\; \widetilde{H}\Big(P, q + \frac{\partial \widetilde{S}}{\partial P}(P,q,t)\Big), \qquad \widetilde{S}(P,q,0) = 0. \tag{3.5}$$

Since $\widetilde{H}$ depends on the parameter h, this is also the case for $\widetilde{S}$. Our aim is to determine the functions $H_j(p,q)$ such that the solution $\widetilde{S}(P,q,t)$ of (3.5) coincides for $t = h$ with (3.3).

We first express $\widetilde{S}(P,q,t)$ as a series

$$\widetilde{S}(P,q,t) \;=\; t\,\widetilde{S}_1(P,q,h) + t^2 \widetilde{S}_2(P,q,h) + t^3 \widetilde{S}_3(P,q,h) + \ldots\;,$$

insert it into (3.5) and compare powers of t. This allows us to obtain the functions $\widetilde{S}_j(p,q,h)$ recursively in terms of derivatives of $\widetilde{H}$:

$$\begin{aligned}
\widetilde{S}_1(p,q,h) &\;=\; \widetilde{H}(p,q) \\
2\,\widetilde{S}_2(p,q,h) &\;=\; \Big(\frac{\partial \widetilde{H}}{\partial q}\cdot\frac{\partial \widetilde{S}_1}{\partial P}\Big)(p,q,h) \\
3\,\widetilde{S}_3(p,q,h) &\;=\; \Big(\frac{\partial \widetilde{H}}{\partial q}\cdot\frac{\partial \widetilde{S}_2}{\partial P}\Big)(p,q,h) + \frac{1}{2}\Big(\frac{\partial^2 \widetilde{H}}{\partial q^2}\Big(\frac{\partial \widetilde{S}_1}{\partial P},\frac{\partial \widetilde{S}_1}{\partial P}\Big)\Big)(p,q,h).
\end{aligned} \tag{3.6}$$

We then write $\widetilde{S}_j$ as a series

$$\widetilde{S}_j(p,q,h) = \widetilde{S}_{j1}(p,q) + h\,\widetilde{S}_{j2}(p,q) + h^2 \widetilde{S}_{j3}(p,q) + \ldots\;,$$

insert it and the expansion (3.4) for $\widetilde{H}$ into (3.6), and compare powers of h. This yields $\widetilde{S}_{1k}(p,q) = H_k(p,q)$ and for $j > 1$ we see that $\widetilde{S}_{jk}(p,q)$ is a function of derivatives of H_l with $l < k$.

The requirement $S(p,q,h) = \widetilde{S}(p,q,h)$ finally shows $S_1(p,q) = \widetilde{S}_{11}(p,q)$, $S_2(p,q) = \widetilde{S}_{12}(p,q) + \widetilde{S}_{21}(p,q)$, etc., so that

$$S_j(p,q) = H_j(p,q) + \text{“function of derivatives of } H_k(p,q) \text{ with } k < j\text{”}.$$

For a given generating function $S(P,q,h)$, this recurrence relation allows us to determine successively the $H_j(p,q)$. We see from these explicit formulas that the functions H_j are defined on the same domain as the S_j. □

As a consequence of Theorem 3.2 and Theorems VI.5.4 and VI.5.5 we obtain the following result.

Theorem 3.3. *A symplectic (partitioned) Runge–Kutta method applied to a system with smooth Hamiltonian $H : D \to \mathbb{R}$ (with $D \subset \mathbb{R}^{2d}$ an arbitrary open set) has a modified Hamiltonian (3.4) with smooth functions $H_j : D \to \mathbb{R}$.* □

Example 3.4 (Symplectic Euler Method). The symplectic Euler method is nothing other than (3.2) with $S(P,q,h) = h\,H(P,q)$. We therefore have (3.3) with $S_1(p,q) = H(p,q)$ and $S_j(p,q) = 0$ for $j > 1$. Following the constructive proof of Theorem 3.2 we obtain

$$\widetilde{H} = H - \frac{h}{2} H_p H_q + \frac{h^2}{12}\Big(H_{pp} H_q^2 + H_{qq} H_p^2 + 4 H_{pq} H_q H_p\Big) + \ldots \qquad (3.7)$$

as the modified Hamiltonian of the symplectic Euler method. For vector-valued p and q, the expression $H_p H_q$ is the scalar product of the vectors H_p and H_q, and $H_{pp} H_q^2 = H_{pp}(H_q, H_q)$ with the second derivative interpreted as a bilinear mapping.

As a particular example consider the pendulum problem (I.1.13), which is Hamiltonian with $H(p,q) = p^2/2 - \cos q$, and apply the symplectic Euler method. By (3.7), the modified Hamiltonian is

$$\widetilde{H}(p,q) = H(p,q) - \frac{h}{2}\, p \sin q + \frac{h^2}{12}\Big(\sin^2 q + p^2 \cos q\Big) + \ldots .$$

This example illustrates that the modified equation corresponding to a separable Hamiltonian (i.e., $H(p,q) = T(p) + U(q)$) is in general not separable. Moreover, it shows that the modified equation of a second order differential equation $\ddot{q} = -\nabla U(q)$ (or equivalently, $\dot{q} = p, \dot{p} = -\nabla U(q)$) is in general not a second order equation.

In principle, the constructive proof of Theorem 3.2 allows us to explicitly compute the modified equation of every symplectic (partitioned) Runge–Kutta method. In Sect. IX.9.3 below we shall, however, give explicit formulas for the modified Hamiltonian in terms of trees. This also yields an alternative proof of Theorem 3.3.

IX.3.3 Poisson Integrators

Consider a Poisson system, i.e., a differential equation

$$\dot{y} = B(y)\nabla H(y), \tag{3.8}$$

where the structure matrix $B(y)$ satisfies the conditions of Lemma VII.2.3, and apply a Poisson integrator (Definition VII.4.6).

Theorem 3.5. *If a Poisson integrator $\Phi_h(y)$ is applied to the Poisson system (3.8), then the modified equation is locally a Poisson system. More precisely, for every $y_0 \in \mathbb{R}^n$ there exist a neighbourhood U and smooth functions $H_j : U \to \mathbb{R}$ such that on U, the modified equation is of the form*

$$\dot{\widetilde{y}} \,=\, B(\widetilde{y})\Big(\nabla H(\widetilde{y}) + h\,\nabla H_2(\widetilde{y}) + h^2\nabla H_3(\widetilde{y}) + \ldots\Big). \tag{3.9}$$

Proof. We use the local change of coordinates $(u,c) = \chi(y)$ of the Darboux–Lie Theorem. By Corollary VII.3.6, this transforms (3.8) to

$$\dot{u} = J^{-1}\nabla_u K(u,c), \qquad \dot{c} = 0,$$

where $K(u,c) = H(y)$ and ∇_u is the gradient with respect to u. The same transformation takes $\Phi_h(y)$ to $\chi\circ\Phi_h\circ\chi^{-1}(u,c) = \big(\Psi_h^1(u,c),c\big)$, where by Lemma VII.4.10 $u \mapsto \Psi_h^1(u,c)$ is a symplectic transformation for every c. By Theorem 3.1, the modified equation in the (u,c) variables is of the form

$$\dot{\widetilde{u}} = J^{-1}\nabla_u \widetilde{K}(\widetilde{u},\widetilde{c}), \qquad \dot{\widetilde{c}} = 0$$

with $\widetilde{K}(u,c) = K(u,c) + h\,K_2(u,c) + h^2 K_3(u,c) + \ldots$. Transforming back to the y-variables gives the modified equation (3.9) with $H_j(y) = K_j(u,c)$. □

The above result is purely local in that it relies on the local transformation of the Darboux–Lie Theorem. It can be made more global under additional conditions on the differential equation.

Theorem 3.6. *If $H(y)$ and $B(y)$ are defined and smooth on a simply connected domain D, and if $B(y)$ is invertible on D, then a Poisson integrator $\Phi_h(y)$ has a modified equation (3.9) with smooth functions $H_j(y)$ defined on all of D.*

Proof. By the construction of Sect. IX.1, the coefficient functions $f_j(y)$ of the modified equation (1.1) are defined and smooth on D. Since $B(y)$ is assumed invertible, there exist unique smooth functions $g_j(y)$ such that $f_j(y) = B(y)g_j(y)$. It remains to show that $g_j(y) = \nabla H_j(y)$ for a function $H_j(y)$ defined on D.

By the local result of Theorem 3.5, we know that for every $y_0 \in D$ there exist functions $H_j^0(y)$ such that $g_j(y) = \nabla H_j^0(y)$ in a neighbourhood of y_0. This implies that the Jacobian of $g_j(y)$ is symmetric on D. The Integrability Lemma VI.2.7 thus proves the existence of functions $H_j(y)$ defined on all of D such that $g_j(y) = \nabla H_j(y)$. □

IX.4 Modified Equations of Splitting Methods

For splitting methods applied to a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y), \tag{4.1}$$

the modified differential equation is obtained directly with the calculus of Lie derivatives and the Baker-Campbell-Hausdorff formula. This approach is due to Yoshida (1993) who considered the case of separable Hamiltonian systems.

First-Order Splitting. Consider the splitting method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]},$$

where $\varphi_h^{[i]}$ is the time-h flow of $\dot{y} = f^{[i]}(y)$. In terms of the Lie derivatives D_i defined by $D_i g(y) = g'(y) f^{[i]}(y)$, this method becomes, using Lemma III.5.1,

$$\Phi_h = \exp(hD_2)\exp(hD_1)\mathrm{Id},$$

and with the BCH formula (III.4.11), (III.4.12) this reads

$$\Phi_h = \exp(h\widetilde{D})\mathrm{Id}$$

with

$$\widetilde{D} = D_1 + D_2 + \frac{h}{2}\,[D_2, D_1] + \frac{h^2}{12}\Big([D_2,[D_2,D_1]] + [D_1,[D_1,D_2]]\Big) + \ldots . \tag{4.2}$$

It follows that Φ_h is formally the exact time-h flow of the modified equation

$$\dot{\widetilde{y}} = \widetilde{f}(\widetilde{y}) \qquad \text{with} \qquad \widetilde{f} = \widetilde{D}\,\mathrm{Id}. \tag{4.3}$$

This gives

$$\widetilde{f}(y) = f(y) + hf_2(y) + h^2 f_3(y) + \ldots$$

with $f = f^{[1]} + f^{[2]}$ and

$$\begin{aligned} f_2 &= \frac{1}{2}\Big(f^{[1]\prime} f^{[2]} - f^{[2]\prime} f^{[1]}\Big) \\ f_3 &= \frac{1}{12}\Big(f^{[1]\prime\prime}(f^{[2]}, f^{[2]}) + f^{[1]\prime} f^{[2]\prime} f^{[2]} - f^{[2]\prime\prime}(f^{[1]}, f^{[2]}) - f^{[2]\prime} f^{[1]\prime} f^{[2]} \\ &\qquad + f^{[2]\prime\prime}(f^{[1]}, f^{[1]}) + f^{[2]\prime} f^{[1]\prime} f^{[1]} - f^{[1]\prime\prime}(f^{[2]}, f^{[1]}) - f^{[1]\prime} f^{[2]\prime} f^{[1]}\Big). \end{aligned}$$

Strang Splitting. For the symmetric splitting

$$\Phi_h^{[\mathrm{S}]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$$

the symmetric BCH formula (III.4.14), (III.4.15) yields

$$\Phi_h^{[S]} = \exp(\frac{h}{2}D_1)\exp(hD_2)\exp(\frac{h}{2}D_1)\,\mathrm{Id} = \exp(h\widetilde{D}^{[S]})\,\mathrm{Id}$$

with

$$\widetilde{D}^{[S]} = D_1 + D_2 + h^2\Big(-\frac{1}{24}[D_1,[D_1,D_2]] + \frac{1}{12}[D_2,[D_2,D_1]]\Big) + \ldots . \quad (4.4)$$

Hence, $\Phi_h^{[S]}$ is the formally exact flow of the modified equation

$$\dot{\widetilde{y}} = \widetilde{f}^{[S]}(\widetilde{y}) \qquad \text{with} \qquad \widetilde{f}^{[S]} = \widetilde{D}^{[S]}\,\mathrm{Id}. \quad (4.5)$$

This gives

$$\widetilde{f}^{[S]}(y) = f(y) + h^2 f_3^{[S]}(y) + h^4 f_5^{[S]}(y) + \ldots$$

with $f = f^{[1]} + f^{[2]}$ and

$$\begin{aligned} f_3^{[S]} \;=\; & \Big(\frac{1}{12}\big(f^{[1]\prime\prime}(f^{[2]},f^{[2]}) + f^{[1]\prime}f^{[2]\prime}f^{[2]} - f^{[2]\prime\prime}(f^{[1]},f^{[2]}) - f^{[2]\prime}f^{[1]\prime}f^{[2]}\big) \\ & -\frac{1}{24}\big(f^{[2]\prime\prime}(f^{[1]},f^{[1]}) + f^{[2]\prime}f^{[1]\prime}f^{[1]} - f^{[1]\prime\prime}(f^{[2]},f^{[1]}) - f^{[1]\prime}f^{[2]\prime}f^{[1]}\big)\Big). \end{aligned}$$

The modified equations for general splitting methods (III.5.13) are obtained in the same way, using Lemma III.5.5.

Hamiltonian Splittings. Consider a differential equation (4.1) where the vector fields $f^{[i]}(y) = J^{-1}\nabla H^{[i]}(y)$ are Hamiltonian. Lemma VII.3.1 shows that the commutator of the Lie derivatives of two Hamiltonian vector fields is the Lie derivative of another Hamiltonian vector field which corresponds to the Poisson bracket of the two Hamiltonians: $[D_F, D_G] = D_{\{G,F\}}$. This implies in particular that the modified differential equations (4.3) and (4.5) are again Hamiltonian. For the first-order splitting, we thus get $f_j(y) = J^{-1}\nabla H_j(y)$, where by (4.2) and (4.3),

$$\begin{aligned} H_2 &= \frac{1}{2}\{H^{[1]}, H^{[2]}\} \\ H_3 &= \frac{1}{12}\Big(\{\{H^{[1]},H^{[2]}\},H^{[2]}\} + \{\{H^{[2]},H^{[1]}\},H^{[1]}\}\Big), \end{aligned}$$

and for the Strang splitting, by (4.4) and (4.5),

$$H_3^{[S]} = -\frac{1}{24}\{\{H^{[2]},H^{[1]}\},H^{[1]}\} + \frac{1}{12}\{\{H^{[1]},H^{[2]}\},H^{[2]}\}.$$

The explicit expressions from the BCH-formula show that the modified Hamiltonian is defined on the same open set as the smooth Hamiltonians $H^{[i]}$.

For the splitting $H(p,q) = T(p) + U(q)$ of a separable Hamiltonian, this approach gives an alternative derivation of the modified equation (3.7) of the symplectic Euler method, and a simple construction of the modified equation of the Störmer–Verlet method (Yoshida 1993). Here, the formula simplifies to

$$\widetilde{H}^{[S]} \;=\; H + h^2\Big(-\frac{1}{24}U_{qq}(T_p,T_p) + \frac{1}{12}T_{pp}(U_q,U_q)\Big) + \ldots . \quad (4.6)$$

IX.5 Modified Equations of Methods on Manifolds

We consider the relationship between numerical methods for differential equations on manifolds and the associated modified differential equations. We give applications to the study of first integrals, constrained Hamiltonian systems, and Lie–Poisson integrators.

IX.5.1 Methods on Manifolds and First Integrals

Consider a differential equation on a smooth manifold $\mathcal{M}$,

$$\dot{y} = f(y) \qquad \text{with} \qquad f(y) \in T_y\mathcal{M}, \tag{5.1}$$

with a smooth vector field $f(y)$ defined on $\mathcal{M}$.

Theorem 5.1. *Let $\Phi_h : \mathcal{M} \to \mathcal{M}$ be an integrator on the manifold $\mathcal{M}$, with $\Phi_h(y)$ depending smoothly on (y, h). Then, there exists a modified differential equation on $\mathcal{M}$,*

$$\dot{\widetilde{y}} = f(\widetilde{y}\,) + hf_2(\widetilde{y}\,) + h^2 f_3(\widetilde{y}\,) + \ldots \tag{5.2}$$

with smooth $f_j(y) \in T_y\mathcal{M}$, such that $\varphi_{r,h}(y) = \Phi_h(y) + \mathcal{O}(h^{r+1})$, where $\varphi_{r,t}(y)$ denotes the flow of the truncation of (5.2) after r terms.

For symmetric methods, the expansion (5.2) contains only even powers of h.

Proof. We choose a local parametrization $y = \chi(z)$ of the manifold $\mathcal{M}$. In the coordinates z the differential equation (5.1) reads

$$\dot{z} = F(z) \quad \text{with } F(z) \text{ defined by} \quad \chi'(z)F(z) = f(\chi(z)),$$

and the numerical integrator becomes

$$\Psi_h(z) = \chi^{-1} \circ \Phi_h \circ \chi(z).$$

Since $F(z)$ and $\Psi_h(z)$ are smooth, the standard backward error analysis on $\mathbb{R}^n$ of Sect. IX.1 yields a modified equation for the integrator $\Psi_h(z)$,

$$\dot{\widetilde{z}} = F(\widetilde{z}\,) + hF_2(\widetilde{z}\,) + h^2 F_3(\widetilde{z}\,) + \ldots \,.$$

Defining

$$f_j(y) = \chi'(z)\, F_j(z) \qquad \text{for} \quad y = \chi(z)$$

gives the desired vector fields $f_j(y)$ on $\mathcal{M}$. It follows from the uniqueness of the modified equation in the parameter space that $f_j(y)$ is independent of the choice of the local parametrization.

The additional statement on symmetric methods follows from Theorem 2.2, because Ψ_h is symmetric if and only if Φ_h is symmetric. □

Under an analyticity assumption, the converse statement also holds.

Theorem 5.2. *Let the integrator $\Phi_h : U \to \mathbb{R}^n$ (with open $U \subset \mathbb{R}^n$) be real analytic in h, and let $\mathcal{M} = \{y \in U\,;\, g(y) = 0\}$ with real analytic $g : U \to \mathbb{R}^m$. If the coefficient functions $f_j(y)$ of the modified differential equation (5.2) satisfy $g'(y)f_j(y) = 0$ for all j and all $y \in \mathcal{M}$, then the restriction of Φ_h to $\mathcal{M}$ defines an integrator on $\mathcal{M}$, i.e., $\Phi_h : \mathcal{M} \to \mathcal{M}$.*

Proof. By the assumption on $f_j(y)$, the flow of the truncated modified equation satisfies $g \circ \varphi_{r,h}(y) = 0$ for all $r \geq 1$ and all $y \in \mathcal{M}$. Since $\varphi_{r,h}(y) = \Phi_h(y) + \mathcal{O}(h^{r+1})$, we have $g \circ \Phi_h(y) = \mathcal{O}(h^{r+1})$ for all r. The analyticity assumptions therefore imply $g \circ \Phi_h(y) = 0$. □

Theorems 5.1 and 5.2 apply to many situations treated in Chap. IV.

First Integrals. The following result was obtained by Gonzalez, Higham & Stuart (1999) and Reich (1999) with different arguments.

Corollary 5.3. *Consider a differential equation $\dot{y} = f(y)$ with a first integral $I(y)$, i.e., $I'(y)f(y) = 0$ for all y. If the numerical method preserves this first integral, then every truncation of the modified equation has $I(y)$ as a first integral.*

Proof. This follows from Theorem 5.1 by considering $\dot{y} = f(y)$ as a differential equation on the manifold $\mathcal{M} = \{y\,;\, I(y) = Const\}$, for which the tangent space is $T_y\mathcal{M} = \{v\,;\, I'(y)v = 0\}$. □

The following converse of Corollary 5.3 is a direct consequence of Theorem 5.2.

Corollary 5.4. *Consider a differential equation $\dot{y} = f(y)$ with a real-analytic first integral $I(y)$. If the numerical method $\Phi_h(y)$ is real analytic in h, and if every truncation of the modified equation has $I(y)$ as a first integral, then the numerical method preserves $I(y)$ exactly, i.e., $I\big(\Phi_h(y)\big) = I(y)$ for all y.* □

Projection Methods. Algorithm IV.4.2 defines a smooth mapping on the manifold if the direction of projection depends smoothly on the position. This is satisfied by orthogonal projection, but is not fulfilled if switching coordinate projections are used (as in Example 4.3). The symmetric orthogonal projection method of Algorithm V.4.1 gives a symmetric method on the manifold to which Theorem 5.1 can be applied.

Methods Based on Local Coordinates. If the parametrization of the manifold employed in Algorithms IV.5.3 and V.4.5 depends smoothly on the position, then again Theorem 5.1 applies. This is the case for the tangent space parametrization, but not for the generalized coordinate partitioning considered in Sect. IV.5.3.

Corollary 5.5 (Lie Group Methods). *Consider a differential equation on a matrix Lie group G,*

$$\dot{Y} = A(Y)Y,$$

where $A(Y)$ is in the associated Lie algebra $\mathfrak{g}$. A Lie group integrator $\Phi_h : G \to G$ has the modified equation

$$\dot{\widetilde{Y}} = \big(A(\widetilde{Y}) + hA_2(\widetilde{Y}) + h^2 A_3(\widetilde{Y}) + \ldots\big)\widetilde{Y} \tag{5.3}$$

with $A_j(Y) \in \mathfrak{g}$ *for* $Y \in G$.

Proof. This is a direct consequence of Theorem 5.1 and (IV.6.3), viz., $T_Y G = \{AY | A \in \mathfrak{g}\}$. □

IX.5.2 Constrained Hamiltonian Systems

In Sect. VII.1 we studied symplectic numerical integrators for constrained Hamiltonian systems

$$\begin{aligned} \dot q &= H_p(p,q) \\ \dot p &= -H_q(p,q) - G(q)^T\lambda \\ 0 &= g(q). \end{aligned} \tag{5.4}$$

Assuming the regularity condition (VII.1.13), the Lagrange parameter $\lambda = \lambda(p,q)$ is given by (VII.1.12). This system can be interpreted as a differential equation on the manifold

$$\mathcal{M} = \big\{(p,q) \mid g(q) = 0,\ G(q)H_p(p,q) = 0\big\}, \tag{5.5}$$

where $G(q) = g'(q)$. The symplectic Euler method (VII.1.19)–(VII.1.20), the RATTLE scheme (VII.1.26), and the Lobatto IIIA-IIIB pair (VII.1.27)–(VII.1.30) were found to be symplectic integrators Φ_h on the manifold $\mathcal{M}$.

Theorem 5.6. *A symplectic integrator $\Phi_h : \mathcal{M} \to \mathcal{M}$ for the constrained Hamiltonian system (5.4) has a modified equation which is locally of the form*

$$\begin{aligned} \dot{\widetilde q} &= \widetilde H_p(\widetilde p,\widetilde q) \\ \dot{\widetilde p} &= -\widetilde H_q(\widetilde p,\widetilde q) - G(\widetilde q)^T\widetilde\lambda \\ 0 &= g(\widetilde q), \end{aligned} \tag{5.6}$$

where $\widetilde\lambda = \widetilde\lambda(\widetilde p,\widetilde q)$ is given by (VII.1.12) with H replaced by $\widetilde H$, and

$$\widetilde H(p,q) = H(p,q) + h\,H_2(p,q) + h^2 H_3(p,q) + \ldots \tag{5.7}$$

with $H_j(p,q)$ satisfying $G(q)\nabla_p H_j(p,q) = 0$ for $(p,q) \in \mathcal{M}$ and all j.

Proof. As explained in Example VII.2.7, a local parametrization $(p,q) = \chi(z)$ of the manifold $\mathcal{M}$ transforms (5.4) to the Poisson system

$$\dot z = B(z)\nabla K(z) \tag{5.8}$$

with $B(z) = (\chi'(z)^T J \chi'(z))^{-1}$ and $K(z) = H(\chi(z))$. Lemma VII.4.9 implies that the numerical method $\Phi_h(p,q)$ on $\mathcal{M}$ becomes a Poisson integrator $\Psi_h(z)$ for (5.8). By Theorem 3.5, $\Psi_h(z)$ has the modified equation

$$\dot{\widetilde z} = B(\widetilde z)\Big(\nabla K(\widetilde z) + h\,\nabla K_2(\widetilde z) + h^2\nabla K_3(\widetilde z) + \ldots\Big). \tag{5.9}$$

Let π be a smooth projection onto the manifold $\mathcal{M}$, defined on a neighbourhood of $\mathcal{M}$ in $\mathbb{R}^{2d}$. We then define

$$H_j(p,q) = K_j\big(\chi^{-1}(\pi(p,q))\big) + \mu(p,q)^T G(q)\nabla_p H(p,q)$$

where we choose $\mu(p,q)$ such that

$$G(q)\nabla_p H_j(p,q) = 0 \qquad \text{for} \quad (p,q)\in\mathcal{M}. \tag{5.10}$$

This is possible because of the regularity assumption (VII.1.13), and because $G(q)\nabla_p H(p,q) = 0$ on $\mathcal{M}$. The condition (5.10) implies that the system (5.6) can be viewed as a differential equation on the original manifold $\mathcal{M}$. Using the same parametrization $(p,q) = \chi(z)$ as before shows that (5.6) is equivalent to (5.9). □

We note that, due to the arbitrary choice of the projection π, the functions $H_j(p,q)$ of the modified equation are uniquely defined only on $\mathcal{M}$.

Global Modified Hamiltonian. If we restrict our considerations to partitioned Runge–Kutta methods, it is possible to find $H_j(p,q)$ in (5.7) that are globally defined on $\mathcal{M}$. Such a result is proved by Reich (1996a) and by Hairer & Wanner (1996) for the constrained symplectic Euler method and the RATTLE algorithm, and by Hairer (2003) for general symplectic partitioned Runge–Kutta schemes. We follow the approach of the latter publication, but present the result only for the important special case of the RATTLE algorithm (VII.1.26). The construction of the $H_j(p,q)$ is done in the following three steps.

Step 1. Symplectic Extension of the Method to a Neighbourhood of the Manifold. The numerical solution (p_1,q_1) of (VII.1.26) is well-defined only for initial values satisfying $(p_0,q_0)\in\mathcal{M}$. However, if we replace the condition "$g(q_1)=0$" by

$$g(q_1) = g(q_0) + h\,G(q_0)H_p(p_0,q_0), \tag{5.11}$$

and the condition "$G(q_1)H_p(p_1,q_1)=0$" by

$$G(q_1)H_p(p_1,q_1) = G(q_0)H_p(p_0,q_0), \tag{5.12}$$

then the numerical solution is well-defined for all (p_0,q_0) in an h-independent open neighbourhood of $\mathcal{M}$ (cf. the existence and uniqueness proof of Sect. VII.1.3). Unfortunately, the so-obtained extension of (VII.1.26) is not symplectic.

Inspired by the formula of Lasagni for the generating function of (unconstrained) symplectic Runge–Kutta methods (see Sect. VI.5.2), we let

$$\begin{aligned} S(p_1,q_0,h) &= \frac{h}{2}\Big(H(p_{1/2},q_0) + H(p_{1/2},q_1) + g(q_0)^T\lambda + g(q_1)^T\mu\Big) \\ &- \frac{h^2}{4}\Big(H_q(p_{1/2},q_1) + G(q_1)^T\mu\Big)^T\Big(H_p(p_{1/2},q_0) + H_p(p_{1/2},q_1)\Big), \end{aligned} \tag{5.13}$$

where $p_0, p_{1/2}, p_1, q_0, q_1, \lambda, \mu$ are the values of the above extension. In the definition (5.13) of the generating function we consider $p_0, p_{1/2}, q_1, \lambda, \mu$ as functions of

(p_1, q_0), what is possible because $p_1 = p_0 + \mathcal{O}(h)$. With the help of $S(p, q, h)$ we define a new numerical method on a neighbourhood of $\mathcal{M}$ by

$$p_0 = p_1 + S_q(p_1, q_0, h), \qquad q_1 = q_0 + S_p(p_1, q_0, h). \tag{5.14}$$

This method is symplectic by definition, and it also coincides with the RATTLE algorithm on the manifold $\mathcal{M}$. Using the fact that the last expression in (5.13) equals $(p_1 - p_{1/2})^T(q_1 - q_0)$, this is seen by the same computation as in the proof of Theorem VI.5.4.

Step 2. Application of the Results of Sect. IX.3.2. The function $S(p_1, q_0, h)$ of (5.13) can be expanded into powers of h with coefficients depending on (p_1, q_0). These coefficient functions are composed of derivatives of $H(p, q)$ and $g(q)$ and, consequently, they are globally defined. For example, the h-coefficient is

$$S_1(p_1, q_0) = H(p_1, q_0) + g(q_0)^T \lambda(p_1, q_0), \tag{5.15}$$

where $\lambda(p, q)$ is the function defined in (VII.1.12).

We are thus exactly in the situation, where we can apply Theorem 3.2. This proves that the method (5.14) has a modified differential equation with globally defined modified Hamiltonian

$$\widetilde{H}_{ext}(p, q) = H_1(p, q) + hH_2(p, q) + \ldots . \tag{5.16}$$

In particular, the constructive proof of Theorem 3.2 shows that $H_1(p, q) = S_1(p, q)$ with $S_1(p, q)$ from (5.15).

Step 3. Backinterpretation for the Method on the Manifold. Since the RATTLE algorithm defines a one-step method on $\mathcal{M}$, it follows from Theorem 5.1 that every truncation of the modified differential equation

$$\dot{\widetilde{p}} = -\nabla_q \widetilde{H}_{ext}(\widetilde{p}, \widetilde{q}), \qquad \dot{\widetilde{q}} = \nabla_p \widetilde{H}_{ext}(\widetilde{p}, \widetilde{q}) \tag{5.17}$$

is a differential equation on the manifold $\mathcal{M}$. Terms of the form $g(q)^T \mu(p, q)$ in $\widetilde{H}_{ext}(p, q)$, which vanish on $\mathcal{M}$, give rise to $-g(q)^T \mu_q(p, q) - G(q)^T \mu(p, q)$ and $g(q)^T \mu_p(p, q)$ in the vector field of (5.17). On the manifold $\mathcal{M}$, where $g(q) = 0$, only the expression $-G(q)^T \mu(p, q)$ remains. Consequently, we can arbitrarily remove terms of the form $g(q)^T \mu(p, q)$ from the functions $H_j(p, q)$ in (5.16), if we add a term $-G(q)^T \lambda$ in the differential equation for p with λ defined by the relation $g(q) = 0$. This then gives a problem of the form (5.6) with globally defined $H_j(p, q)$.

IX.5.3 Lie–Poisson Integrators

As in Sect. VII.5.5 we consider a symplectic integrator

$$(P_1, Q_1) = \Phi_h(P_0, Q_0) \qquad \text{on } T^*G$$

for the left-invariant Hamiltonian system (VII.5.43) on a matrix Lie group G with a Hamiltonian $H(P,Q)$ that is quadratic in P. We suppose that the method preserves the left-invariance (VII.5.54) so that it induces a one-step map

$$Y_1 = \Psi_h(Y_0) \qquad \text{on } \mathfrak{g}^*$$

by setting $Y_1 = Q_1^T P_1$ for $(P_1,Q_1) = \Phi_h(P_0,Q_0)$ with $Q_0^T P_0 = Y_0$. This is a numerical integrator for the differential equation (VII.5.37) on $\mathfrak{g}^*$, and in the coordinates $y = (y_j)$ with respect to the basis (F_j) of $\mathfrak{g}^*$ this gives a map

$$y_1 = \psi_h(y_0) \qquad \text{on } \mathbb{R}^d,$$

which is a numerical integrator for the Lie–Poisson system $\dot y = B(y)\nabla H(y)$ with $B(y)$ given by (VII.5.35).

Theorem 5.7. *If $\Phi_h(P,Q)$ is a symplectic and left-invariant integrator for (VII.5.43) which is real analytic in h, then its reduction $\psi_h(y)$ is a Poisson integrator. Moreover, $\Psi_h(Y)$ preserves the coadjoint orbits, i.e., $\Psi_h(Y) \in \{\mathrm{Ad}^*_{U^{-1}} Y \,;\, U \in G\}$.*

Proof. (a) In the first step one shows, by the standard induction argument as in the proof of Theorem 2.3, that the modified equation given by Theorem 5.6,

$$\begin{aligned} \dot{\widetilde P} &= -\nabla_Q \widetilde H(\widetilde P,\widetilde Q) - \sum_{i=1}^m \widetilde\lambda_i \nabla_Q g_i(\widetilde Q), \qquad \dot{\widetilde Q} = \nabla_P \widetilde H(\widetilde P,\widetilde Q) \\ 0 &= g_i(\widetilde Q), \quad i = 1,\ldots,m, \end{aligned} \tag{5.18}$$

with

$$\widetilde H(P,Q) = H(P,Q) + hH_2(P,Q) + h^2 H_3(P,Q) + \ldots$$

is left-invariant, i.e.,

$$H_j(U^T P, U^{-1}Q) = H_j(P,Q) \qquad \text{for all } U \in G \text{ and all } j. \tag{5.19}$$

(b) The Lie–Poisson reduction of Theorem VII.5.8 yields that if $\big(\widetilde P(t),\widetilde Q(t)\big) \in T^*G$ is a solution of the modified system (5.18), then $\widetilde Y(t) = \widetilde Q(t)^T \widetilde P(t) \in \mathfrak{g}^*$ solves the differential equation

$$\langle \dot{\widetilde Y}, X\rangle = \langle \widetilde Y, [\widetilde H'(\widetilde Y), X]\rangle \qquad \text{for all } X \in \mathfrak{g}. \tag{5.20}$$

Theorem VII.5.6 shows that its solution lies on a coadjoint orbit. By Theorem VII.5.5, (5.20) is equivalent to the Poisson system

$$\dot{\widetilde y} = B(\widetilde y)\nabla \widetilde H(\widetilde y). \tag{5.21}$$

(c) We know already from Theorem VII.5.11 that $\psi_h(y)$ is a Poisson map. Since all truncations of the modified equation (5.21) have the Casimirs as first integrals, their preservation by ψ_h follows from Corollary 5.4. Similarly, the preservation of the coadjoint orbits follows from Theorem 5.2. □

In contrast to Theorem 3.5, we here obtain a global modified Hamiltonian in the modified Poisson system if the method is obtained by the discrete Lie–Poisson reduction of the RATTLE algorithm; see the preceding subsection.

IX.6 Modified Equations for Variable Step Sizes

The modified differential equation of a numerical integrator depends on the step size employed. Therefore, if the step size is changed arbitrarily, a different modified equation occurs at every step. This is the reason for the poor longtime behaviour observed in Sect. VIII.1. On the other hand, a satisfactory backward error analysis is possible for the variable-step approaches of Sects. VIII.2 and VIII.3.

Time Transformations. The adaptive approaches of Sect. VIII.2 amount to applying a fixed step size method to a transformed differential equation. Hence, the backward error analysis considered so far applies directly and yields modified equations for the transformed problem. These modified equations are Hamiltonian for Algorithm VIII.2.1 and reversible for method (VIII.2.12).

Proportional, Reversible Step Size Controllers. As in Sect. VIII.3.1 we let the step size be of the form

$$h_{n+1/2} = \varepsilon\, s(y_n, \varepsilon), \tag{6.1}$$

where ε is a small accuracy parameter. It is not allowed to use information from previous steps. The idea is to work with expansions in powers of the fixed parameter ε instead of the step sizes, and to consider the exact solution of the modified equation on a variable grid. The following development is given in Hairer & Stoffer (1997). It extends the results of Sects. IX.1 and IX.2 to variable step sizes.

Theorem 6.1. *Let $\Phi_h(y)$ be a smooth one-step method.*

a) The variable-step method $y \mapsto \Phi_{\varepsilon\, s(y,\varepsilon)}(y)$ has a modified differential equation

$$\dot{\widetilde{y}} = f(\widetilde{y}) + \varepsilon\, f_2(\widetilde{y}) + \varepsilon^2 f_3(\widetilde{y}) + \ldots\,, \tag{6.2}$$

with smooth vector fields $f_j(y)$, such that

$$\varphi_{r,\varepsilon\, s(y,\varepsilon)}(y) = \Phi_{\varepsilon\, s(y,\varepsilon)}(y) + \mathcal{O}(\varepsilon^{r+1})\,, \tag{6.3}$$

where $\varphi_{r,t}(y)$ denotes the flow of the truncation of (6.2) after r terms.

b) If the method is symmetric (i.e., $\Phi_h(y) = \Phi_{-h}^{-1}(y)$) and $s(\widehat{y}, -\varepsilon) = s(y,\varepsilon)$ holds with $\widehat{y} = \Phi_{\varepsilon s(y,\varepsilon)}(y)$, then the expansion (6.2) is in even powers of ε, i.e.,

$$f_j(y) = 0 \qquad \text{for even } j. \tag{6.4}$$

c) If the method is ρ-reversible (i.e., $\rho \circ \Phi_h = \Phi_h^{-1} \circ \rho$) and $s(\rho^{-1}\widehat{y}, \varepsilon) = s(y,\varepsilon)$ holds with $\widehat{y} = \Phi_{\varepsilon s(y,\varepsilon)}(y)$, then the modified equation (6.2) is ρ-reversible, i.e.,

$$\rho \circ f_j = -f_j \circ \rho \qquad \text{for all } j. \tag{6.5}$$

Proof. a) The modified equation (6.2) is constructed by Taylor expansion of (6.3) in the same way as (1.1), using ε-expansions instead of h-expansions.

For the proof of the statements (b) and (c) we denote, as we did in Sect. VIII.3, $\Psi_\varepsilon(y) = \Phi_{\varepsilon s(y,\varepsilon)}(y)$. We then compute the dominant error term in (6.3) and obtain

$$\Psi_\varepsilon(y) = \varphi_{r,\varepsilon s(y,\varepsilon)}(y) + \varepsilon^{r+1} s(y,\varepsilon) f_{r+1}(y) + \mathcal{O}(\varepsilon^{r+2}). \tag{6.6}$$

With the aim of getting an analogous formula for Ψ_ε^{-1}, we put $\widehat{y} = \Psi_\varepsilon(y)$ and use $\varphi_{r,t}^{-1}(y) = \varphi_{r,-t}(y)$ so that

$$y = \varphi_{r,-\varepsilon s(y,\varepsilon)}\big(\widehat{y} - \varepsilon^{r+1} s(y,\varepsilon) f_{r+1}(y) + \mathcal{O}(\varepsilon^{r+2})\big). \tag{6.7}$$

b) Inserting $s(y,\varepsilon) = s(\widehat{y}, -\varepsilon)$ into (6.7) and using the facts that $y = \widehat{y} + \mathcal{O}(\varepsilon)$ and that the derivative $\varphi_{r,t}'(y)$ is $\mathcal{O}(t)$-close to the identity, we obtain

$$\Psi_\varepsilon^{-1}(\widehat{y}) = y = \varphi_{r,-\varepsilon s(\widehat{y},-\varepsilon)}(\widehat{y}) - \varepsilon^{r+1} s(\widehat{y},0) f_{r+1}(\widehat{y}) + \mathcal{O}(\varepsilon^{r+2}). \tag{6.8}$$

By (VIII.3.3) we have $\Psi_\varepsilon = \Psi_{-\varepsilon}^{-1}$. Changing the sign of ε in (6.8), a comparison with (6.6) proves that $f_{r+1}(y) = (-1)^r f_{r+1}(y)$ implying (6.4).

c) With $s(y,\varepsilon) = s(\rho^{-1}\widehat{y},\varepsilon)$ formula (6.7) yields

$$\Psi_\varepsilon^{-1}(\widehat{y}) = \varphi_{r,-\varepsilon s(\rho^{-1}\widehat{y},\varepsilon)}(\widehat{y}) - \varepsilon^{r+1} s(\rho^{-1}\widehat{y},0) f_{r+1}(\widehat{y}) + \mathcal{O}(\varepsilon^{r+2}).$$

By an induction argument on r we assume that $\rho \circ \varphi_{r,t} = \varphi_{r,-t} \circ \rho$. The ρ-reversibility of Ψ_ε, i.e., $\rho \circ \Psi_\varepsilon = \Psi_\varepsilon^{-1} \circ \rho$, thus implies the statement (6.5). □

Integrating, Reversible Step Size Controllers. We next study a backward error analysis for Algorithm VIII.3.4. It is possible to interpret this algorithm as the fixed step size method $\widehat{\Phi}_\varepsilon$ of (VIII.3.19) applied to the augmented system (VIII.3.17) and to apply the construction of Sect. IX.1. This approach has been taken in Hairer & Söderlind (2004). In view of an error analysis for reversible integrable systems it seems to be more convenient to consider the solution of the modified equation on a variable grid as it is done in Theorem 6.1.

Let us recall Algorithm VIII.3.4. For a given basic integrator $\Phi_h(y)$ and a given time transformation $\sigma(y)$ we denote $G(y) = -\big(\sigma(y)\big)^{-1}\nabla\sigma(y)^T f(y)$ and we compute for a given initial value y_0 and with $z_0 = 1/\sigma(y_0)$

$$\begin{aligned} z_{n+1/2} &= z_n + \varepsilon\, G(y_n)/2 \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \\ z_{n+1} &= z_{n+1/2} + \varepsilon\, G(y_{n+1})/2. \end{aligned} \tag{6.9}$$

The values y_n approximate $y(t_n)$, where $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$. We further use the notation

$$\Psi_\varepsilon : \begin{pmatrix} y_n \\ z_n \end{pmatrix} \mapsto \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} \qquad \text{and} \qquad \widehat{\rho} = \begin{pmatrix} \rho & 0 \\ 0 & 1 \end{pmatrix}. \tag{6.10}$$

The step size used in this algorithm is

$$h_{n+1/2} = \frac{\varepsilon}{z_{n+1/2}} = \varepsilon s(y_n, z_n, \varepsilon) \qquad \text{with} \qquad s(y,z,\varepsilon) = \frac{1}{z + \varepsilon G(y)/2}. \tag{6.11}$$

The symmetric definition of the algorithm immediately yields

$$s(\widehat{y},\widehat{z},-\varepsilon) = s(y,z,\varepsilon) \qquad \text{for} \qquad (\widehat{y},\widehat{z}) = \Psi_\varepsilon(y,z). \tag{6.12}$$

For a ρ-reversible differential equation $\dot{y} = f(y)$ and for $\sigma(y)$ satisfying $\sigma(\rho^{-1}y) = \sigma(y)$ we have $G(\rho^{-1}y) = -G(y)$. Consequently, the step size function $s(y,z,\varepsilon)$ of (6.11) also satisfies

$$s(\rho^{-1}\widehat{y},\widehat{z},-\varepsilon) = s(y,z,\varepsilon) \qquad \text{for} \qquad (\widehat{y},\widehat{z}) = \Psi_\varepsilon(y,z). \tag{6.13}$$

With this preparation we are able to formulate the following result.

Theorem 6.2. *Let $\Phi_h(y)$ be a smooth one-step method, $\sigma(y)$ a smooth time transformation, and $s(y,z,\varepsilon)$ the step size function of (6.11).*

a) For the method Ψ_ε of (6.10) there exists a modified differential equation

$$\begin{aligned} \dot{\widetilde{y}} &= f(\widetilde{y}) + \varepsilon\, f_2(\widetilde{y},\widetilde{z}) + \varepsilon^2 f_3(\widetilde{y},\widetilde{z}) + \ldots \\ \dot{\widetilde{z}} &= \widetilde{z}\, G(\widetilde{y}) + \varepsilon\, G_2(\widetilde{y},\widetilde{z}) + \varepsilon^2 G_3(\widetilde{y},\widetilde{z}) + \ldots\ , \end{aligned} \tag{6.14}$$

with smooth vector fields $f_j(y,z), G_j(y,z)$, such that

$$\varphi_{r,\varepsilon\, s(y,z,\varepsilon)}(y,z) = \Psi_\varepsilon(y,z) + \mathcal{O}(\varepsilon^{r+1})\,, \tag{6.15}$$

where $\varphi_{r,t}(y,z)$ denotes the flow of the truncation of the system (6.14) after r terms.

b) If the basic method is symmetric (i.e., $\Phi_h(y) = \Phi_{-h}^{-1}(y)$) then

$$f_j(y) = 0 \qquad \textit{for even } j. \tag{6.16}$$

c) If the basic method is ρ-reversible (i.e., $\rho\circ\Phi_h = \Phi_h^{-1}\circ\rho$) and $\sigma(\rho^{-1}y) = \sigma(y)$ holds, then the modified equation (6.14) is $\widehat{\rho}$-reversible with $\widehat{\rho}$ given by (6.10), i.e.,

$$\rho f_j(y,z) = -f_j(\rho y,z), \quad G_j(y,z) = -G(\rho y,z) \qquad \textit{for all } j. \tag{6.17}$$

Proof. The proof is the same as for Theorem 6.1 and therefore omitted. Notice that the step size function satisfies (6.12) and (6.13) which are needed in that proof. □

If the basic method is of order p then the coefficient functions of (6.14) satisfy $f_j(y,z) = 0$ for $j = 2,\ldots,p$. We always have $G_2(y,z) = 0$ due to the symmetric way of choosing $z_{n+1/2}$ in (6.9). However, $G_3(y,z) \neq 0$ in general, even if the method Φ_h has an order higher than two.

IX.7 Rigorous Estimates – Local Error

> Wherefore it is highly desirable that it be clearly and rigorously shown why series of this kind, which at first converge very rapidly and then ever more slowly, and at length diverge more and more, nevertheless give a sum close to the true one if not too many terms are taken, and to what degree such a sum can safely be considered as exact.
>
> (a footnote in Gauss' thesis, 1799)

Up to now we have considered the modified equation (1.1) as a formal series without taking care of convergence issues. Here,

- we show that already in very simple situations the modified differential equation does not converge;
- we give bounds on the coefficient functions $f_j(y)$ of the modified equation (1.1), so that an optimal truncation index can be determined;
- we estimate the difference between the numerical solution $y_1 = \Phi_h(y_0)$ and the exact solution $\widetilde{y}(h)$ of the truncated modified equation.

These estimates will be the basis for rigorous statements concerning the long-time behaviour of numerical solutions. The rigorous estimates of the present section have been given in the articles Benettin & Giorgilli (1994), Hairer & Lubich (1997) and Reich (1999). We mainly follow the approach of Benettin & Giorgilli, but we also use ideas of the other two papers.

Example 7.1. We consider the differential equation[1] $\dot{y} = f(t)$, $y(0) = 0$, and we apply the trapezoidal rule $y_1 = h\big(f(0) + f(h)\big)/2$. In this case, the numerical solution has an expansion $\Phi_h(t, y) = y + h\big(f(t) + f(t+h)\big)/2 = y + hf(t) + h^2 f'(t)/2 + h^3 f''(t)/4 + \ldots$, so that the modified equation is necessarily of the form

$$\dot{\widetilde{y}} = f(t) + hb_1 f'(t) + h^2 b_2 f''(t) + h^3 b_3 f'''(t) + \ldots . \tag{7.1}$$

The real coefficients b_k can be computed by putting $f(t) = e^t$. The relation $\Phi_h(t, y) = \widetilde{y}(t+h)$ (with initial value $\widetilde{y}(t) = y$) yields after division by e^t

$$\frac{h}{2}\Big(e^h + 1\Big) = \Big(1 + b_1 h + b_2 h^2 + b_3 h^3 + \ldots\Big)\Big(e^h - 1\Big).$$

This proves that $b_1 = 0$, and $b_k = B_k/k!$, where B_k are the Bernoulli numbers (see for example Hairer & Wanner (1997), Sect. II.10). Since these numbers behave like $B_k/k! \approx Const \cdot (2\pi)^{-k}$ for $k \to \infty$, the series (7.1) diverges for all $h \neq 0$, as soon as the derivatives of $f(t)$ grow like $f^{(k)}(t) \approx k!\, MR^{-k}$. This is typically the case for analytic functions $f(t)$ with finite poles.

It is interesting to remark that the relation $\Phi_h(t, y) = \widetilde{y}(t+h)$ is nothing other than the Euler-MacLaurin summation formula.

As a particular example we choose the function

$$f(t) = \frac{5}{1 + 25t^2}.$$

Figure 7.1 shows the numerical solution and the exact solution of the modified equation truncated at different values of N. For $h = 0.2$, there is an excellent agreement for $N \leq 12$, whereas oscillations begin to appear from $N = 14$ onwards. For the halved step size $h = 0.1$, the oscillations become visible for N twice as large.

[1] Observe that after adding the equation $\dot{t} = 1$, $t(0) = 0$, we get for $Y = (t, y)^T$ the autonomous differential equation $\dot{Y} = F(Y)$ with $F(Y) = (1, f(t))^T$. Hence, all results of this chapter are applicable.

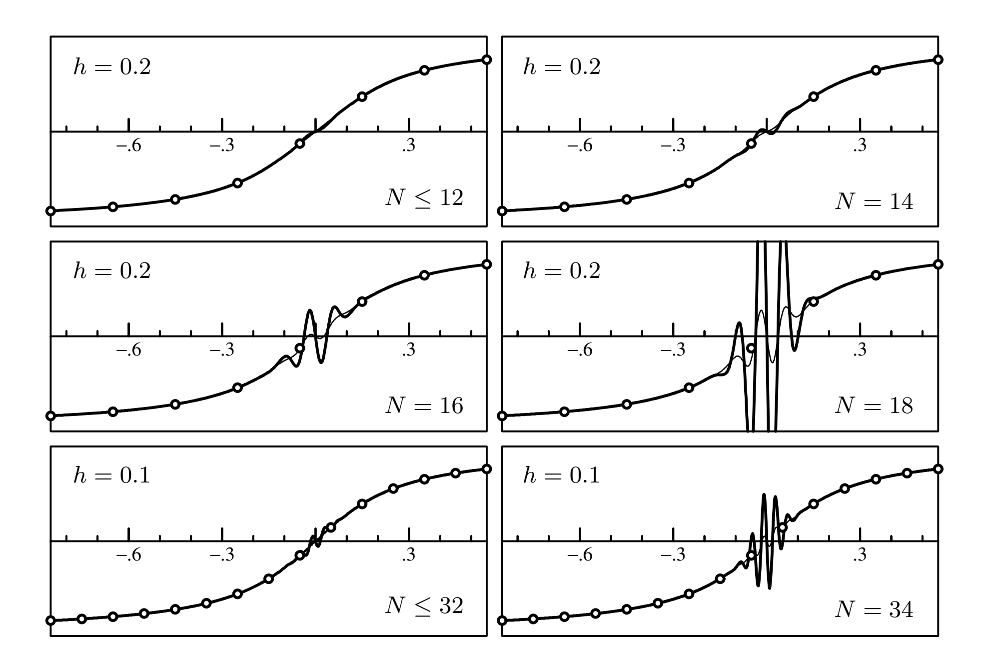


Fig. 7.1. Numerical solution with the trapezoidal rule compared to the solution of the truncated modified equation for $h = 0.2$ (upper four pictures), and for $h = 0.1$ (lower two pictures)

The main ingredient of a rigorous backward error analysis is an analyticity assumption on the differential equation $\dot{y} = f(y)$ and on the method. Throughout this section we assume that $f(y)$ is analytic in a complex neighbourhood of y_0 and that

$$\|f(y)\| \leq M \qquad \text{for} \qquad \|y - y_0\| \leq 2R \tag{7.2}$$

i.e., for all y of $B_{2R}(y_0) := \{y \in \mathbb{C}^d ; \|y - y_0\| \leq 2R\}$. Our strategy is the following: using (7.2) and Cauchy's estimates we derive bounds for the coefficient functions $d_j(y)$ of (1.3) on $B_R(y_0)$ (Sect. IX.7.1), then we estimate the functions $f_j(y)$ of the modified differential equation on $B_{R/2}(y_0)$ (Sect. IX.7.2), and finally we search for a suitable truncation for the formal series (1.1) and we prove the closeness of the numerical solution to the exact solution of the truncated modified equation (Sect. IX.7.3).

IX.7.1 Estimation of the Derivatives of the Numerical Solution

If we apply a numerical method to $\dot{y} = f(y)$ with analytic $f(y)$, the expression $\Phi_h(y)$ will usually be analytic in a neighbourhood of $h = 0$ and $y \in B_R(y_0)$. Consequently, the coefficients $d_j(y)$ of the Taylor series expansion

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \ldots \tag{7.3}$$

are also analytic and the functions $d_j(y)$ can be estimated by the use of Cauchy's inequalities. Let us demonstrate this for Runge–Kutta methods.

Theorem 7.2. *For a Runge–Kutta method (II.1.4) let*

$$\mu = \sum_{i=1}^{s} |b_i|, \qquad \kappa = \max_{i=1,\ldots,s} \sum_{j=1}^{s} |a_{ij}|. \tag{7.4}$$

If $f(y)$ is analytic in the complex ball $B_{2R}(y_0)$ and satisfies (7.2), then the coefficient functions $d_j(y)$ of (7.3) are analytic in $B_R(y_0)$ and satisfy

$$\|d_j(y)\| \le \mu M \left(\frac{2\kappa M}{R}\right)^{j-1} \qquad \textit{for} \qquad \|y - y_0\| \le R. \tag{7.5}$$

Proof. For $y \in B_{3R/2}(y_0)$ and $\|\Delta y\| \le 1$ the function $\alpha(z) = f(y + z\Delta y)$ is analytic for $|z| \le R/2$ and bounded by M. Cauchy's estimate therefore yields

$$\|f'(y)\Delta y\| = \|\alpha'(0)\| \le 2M/R.$$

Consequently, $\|f'(y)\| \le 2M/R$ for $y \in B_{3R/2}(y_0)$ in the operator norm.

For $y \in B_R(y_0)$, the Runge–Kutta method (II.1.4) requires the solution of the nonlinear system $g_i = y + h\sum_{j=1}^{s} a_{ij} f(g_j)$, which can be solved by fixed point iteration. If $|h| 2\kappa M/R \le \gamma < 1$, it represents a contraction on the closed set $\{(g_1, \ldots, g_s)\,;\, \|g_i - y\| \le R/2\}$ and possesses a unique solution. Consequently, the method is analytic for $|h| \le \gamma R/(2\kappa M)$ and $y \in B_R(y_0)$. This implies that the functions $d_j(y)$ of (7.3) are also analytic. Furthermore, $\|\Phi_h(y) - y\| \le |h|\mu M$ for $y \in B_R(y_0)$ so that, again by Cauchy's estimate,

$$\|d_j(y)\| = \frac{1}{j!}\left\|\frac{d^j}{dh^j}\Big(\Phi_h(y) - y\Big)\Big|_{h=0}\right\| \le \mu M \Big(\frac{2\kappa M}{\gamma R}\Big)^{j-1}$$

for $j \ge 1$. The statement is then obtained by considering the limit $\gamma \to 1$. □

Due to the consistency condition $\sum_{i=1}^{s} b_i = 1$, methods with positive weights b_i all satisfy $\mu = 1$. The values μ, κ of some classes of Runge–Kutta methods are given in Table 7.1 (those for the Gauss methods and for the Lobatto IIIA methods have been checked for $s \le 9$ and $s \le 5$, respectively).

Estimates of the type (7.5), possibly with a different interpretation of M and R, hold for all one-step methods which are analytic in h and y, e.g., partitioned Runge–Kutta methods, splitting and composition methods, projection methods, Lie group methods, ...

Table 7.1. The constants μ and κ of formula (7.4)

method	μ	κ	method	μ	κ
explicit Euler	1	0	implicit Euler	1	1
implicit midpoint	1	$1/2$	trapezoidal rule	1	1
Gauss methods	1	c_s	Lobatto IIIA	1	1

IX.7.2 Estimation of the Coefficients of the Modified Equation

At the beginning of this chapter we gave an explicit formula for the first coefficient functions of the modified differential equation (see (1.4)). Using the Lie derivative

$$(D_i g)(y) = g'(y) f_i(y) \tag{7.6}$$

(cf. (VI.5.2)) and $f_1(y) := f(y)$, these formulas can be written as

$$\begin{aligned} f_2(y) &= d_2(y) - \frac{1}{2!}\big(D_1 f_1\big)(y) \\ f_3(y) &= d_3(y) - \frac{1}{3!}(D_1^2 f_1)(y) - \frac{1}{2!}\big(D_2 f_1 + D_1 f_2\big)(y). \end{aligned}$$

We have the following recurrence relation for the general case.

Lemma 7.3. *If the numerical method has an expansion of the form (7.3), then the functions $f_j(y)$ of the modified differential equation (1.1) satisfy*

$$f_j(y) \;=\; d_j(y) \;-\; \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j} \Big(D_{k_1} \ldots D_{k_{i-1}} f_{k_i}\Big)(y),$$

where $k_m \geq 1$ for all m. Observe that the right-hand expression only involves $f_k(y)$ with $k < j$.

Proof. The solution of the modified equation (1.1) with initial value $y(t) = y$ can be formally written as (cf. (1.2))

$$\widetilde{y}(t+h) \;=\; y + \sum_{i\geq 1} \frac{h^i}{i!}\, D^{i-1} F(y),$$

where $F(y) = f_1(y) + h f_2(y) + h^2 f_3(y) + \ldots$ stands for the modified equation, and $hD = hD_1 + h^2 D_2 + h^3 D_3 + \ldots$ for the corresponding Lie operator. We expand the formal sums and obtain

$$\widetilde{y}(t+h) \;=\; y + \sum_{i\geq 1} \frac{1}{i!} \sum_{k_1,\ldots,k_i} h^{k_1+\ldots+k_i} \Big(D_{k_1} \ldots D_{k_{i-1}} f_{k_i}\Big)(y), \tag{7.7}$$

where all $k_m \geq 1$. Comparing like powers of h in (7.3) and (7.7) yields the desired recurrence relations for the functions $f_j(y)$. □

To get bounds for $\|f_j(y)\|$, we have to estimate repeatedly expressions like $\|(D_i g)(y)\|$. The following variant of Cauchy's estimate will be extremely useful.

Lemma 7.4. *For analytic functions $f_i(y)$ and $g(y)$ we have for $0 \leq \sigma < \rho$ the estimate*

$$\|D_i g\|_\sigma \;\leq\; \frac{1}{\rho - \sigma} \cdot \|f_i\|_\sigma \cdot \|g\|_\rho .$$

Here, $\|g\|_\rho := \max\{\|g(y)\|\,;\, y \in B_\rho(y_0)\}$ and $\|f_i\|_\sigma$, $\|D_i g\|_\sigma$ are defined similarly.

Proof. For a fixed $y \in B_\sigma(y_0)$ the function $\alpha(z) = g\big(y + zf_i(y)\big)$ is analytic for $\|z\| \le \varepsilon := (\rho - \sigma)/M$ with $M := \|f_i\|_\sigma$. Since $\alpha'(0) = g'(y)f_i(y) = (D_i g)(y)$, we get from Cauchy's estimate that

$$\|(D_i g)(y)\| = \|\alpha'(0)\| \le \frac{1}{\varepsilon} \sup_{|z|\le\varepsilon} \|\alpha(z)\| \le \frac{M}{\rho-\sigma} \cdot \|g\|_\rho .$$

This proves the statement. □

We are now able to estimate the coefficients $f_j(y)$ of the modified differential equation.

Theorem 7.5. *Let $f(y)$ be analytic in $B_{2R}(y_0)$, let the Taylor series coefficients of the numerical method (7.3) be analytic in $B_R(y_0)$, and assume that (7.2) and (7.5) are satisfied. Then, we have for the coefficients of the modified differential equation*

$$\big\|f_j(y)\big\| \le \ln 2\ \eta\, M \left(\frac{\eta\, M\, j}{R}\right)^{j-1} \qquad \textit{for} \qquad \|y - y_0\| \le R/2, \tag{7.8}$$

where $\eta = 2\max\big(\kappa, \mu/(2\ln 2 - 1)\big)$.

Proof. We fix an index, say J, and we estimate (in the notation of Lemma 7.4)

$$\|f_j\|_{R-(j-1)\delta} \qquad \text{for} \qquad j = 1, 2, \ldots, J,$$

where $\delta = R/(2(J-1))$. This will then lead to the desired estimate for $\|f_J\|_{R/2}$.

In the following we abbreviate $\|\cdot\|_{R-(j-1)\delta}$ by $\|\cdot\|_j$. Using repeatedly Cauchy's estimate of Lemma 7.4 we get for $k_1 + \ldots + k_i = j$ that

$$\begin{aligned} \|D_{k_1} \ldots D_{k_{i-1}} f_{k_i}\|_j &\le \frac{1}{\delta}\, \|f_{k_1}\|_j\, \|D_{k_2} \ldots D_{k_{i-1}} f_{k_i}\|_{j-1} \\ &\le \ \ldots \ \le \frac{1}{\delta^{i-1}}\, \|f_{k_1}\|_j\, \|f_{k_2}\|_{j-1} \cdot \ldots \cdot \|f_{k_i}\|_{j-i+1} \\ &\le \frac{1}{\delta^{i-1}}\, \|f_{k_1}\|_{k_1}\, \|f_{k_2}\|_{k_2} \cdot \ldots \cdot \|f_{k_i}\|_{k_i} . \end{aligned}$$

The last inequality follows from $\|g\|_j \le \|g\|_l$ for $l \le j$, which is an immediate consequence of $B_{R-(j-1)\delta}(y_0) \subset B_{R-(l-1)\delta}(y_0)$. It therefore follows from Lemma 7.3 that

$$\|f_j\|_j \ \le\ \|d_j\|_j + \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j} \frac{1}{\delta^{i-1}}\, \|f_{k_1}\|_{k_1}\, \|f_{k_2}\|_{k_2} \cdot \ldots \cdot \|f_{k_i}\|_{k_i} .$$

By induction on j ($1 \le j \le J$) we obtain that $\|f_j\|_j \le \delta\beta_j$, where β_j is defined by

$$\beta_j \ = \ \frac{\mu M}{\delta} \left(\frac{2\kappa M}{R}\right)^{j-1} + \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j} \beta_{k_1}\, \beta_{k_2} \cdot \ldots \cdot \beta_{k_i} . \tag{7.9}$$

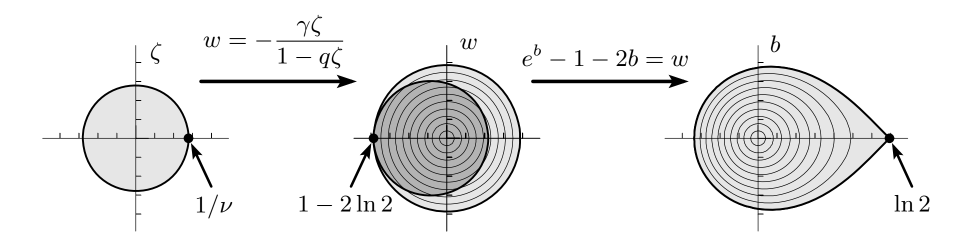


Fig. 7.2. Complex functions of the proof of Theorem 7.5 ($\gamma = q = 1$)

Observe that β_j is defined for all $j \geq 1$. We let $b(\zeta) = \sum_{j\geq 1} \beta_j \zeta^j$ be its generating function and we obtain (by multiplying (7.9) with ζ^j and summing over $j \geq 1$)

$$b(\zeta) = \frac{\gamma\zeta}{1-q\zeta} + \sum_{j\geq 2} \frac{1}{j!}\, b(\zeta)^j = \frac{\gamma\zeta}{1-q\zeta} + e^{b(\zeta)} - 1 - b(\zeta), \tag{7.10}$$

where we have used the abbreviations $\gamma := \mu M/\delta$ and $q := 2\kappa M/R$.

Whenever $e^{b(\zeta)} \neq 2$ (i.e., for $\zeta \neq (2b-1)/\big(\gamma + q(2b-1)\big)$ with $b = \ln 2 + 2k\pi i$) the implicit function theorem can be applied to (7.10). This implies that $b(\zeta)$ is analytic in a disc with radius $1/\nu = (2\ln 2 - 1)/\big(\gamma + q(2\ln 2 - 1)\big)$ and centre at the origin. On the disc $|\zeta| \leq 1/\nu$, the solution $b(\zeta)$ of (7.10) with $b(0) = 0$ is bounded by $\ln 2$. This is seen as follows (Fig. 7.2): with the function $w = -\gamma\zeta/(1 - q\zeta)$ the disc $|\zeta| \leq 1/\nu$ is mapped into a disc which, for all possible choices of $\gamma \geq 0$ and $q \geq 0$, lies in $|w| \leq 2\ln 2 - 1$. The image of this disc under the mapping $b(w)$ defined by $e^b - 1 - 2b = w$ and $b(0) = 0$ is completely contained in the disc $|b| \leq \ln 2$. Cauchy's inequalities therefore imply $|\beta_j| \leq \ln 2 \cdot \nu^j$, and we get

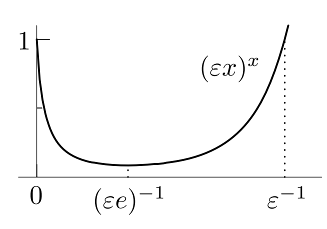
$$\|f_J\|_{R/2} = \|f_J\|_J \leq \delta\beta_J \leq \ln 2 \cdot \delta \cdot \nu^J.$$

Since $\nu = q + \gamma/(2\ln 2 - 1) \leq \eta M J/R$ with η given by $\eta = 2\max\big(\kappa, \mu/(2\ln 2 - 1)\big)$ and $\delta\nu \leq \eta M$, this proves the statement for J. □

IX.7.3 Choice of N and the Estimation of the Local Error

To get rigorous estimates, we truncate the modified differential equation (1.1), and we consider

$$\dot{\widetilde{y}} = F_N(\widetilde{y}\,), \qquad F_N(\widetilde{y}\,) = f(\widetilde{y}\,) + hf_2(\widetilde{y}\,) + \ldots + h^{N-1} f_N(\widetilde{y}\,) \tag{7.11}$$



with initial value $\widetilde{y}(0) = y_0$. It is common in the theory of asymptotic expansions to truncate the series at the index where the corresponding term is minimal. Motivated by the bound (7.8) and by the fact that $(\varepsilon x)^x$ admits a minimum for $x = (\varepsilon e)^{-1}$ (see the picture to the left with $\varepsilon = 0.15$), we suppose that the truncation index N satisfies

$$hN \le h_0 \qquad \text{with} \qquad h_0 = \frac{R}{e\eta M}. \tag{7.12}$$

Under the less restrictive assumption $hN \le eh_0$, the estimates (7.2) and (7.8) imply for $\|y - y_0\| \le R/2$ that

$$\begin{aligned} \|F_N(y)\| \;&\le\; M\Big(1 + \eta\,\ln 2 \sum_{j=2}^{N} \Big(\frac{\eta M j h}{R}\Big)^{j-1}\Big) \\ &\le\; M\Big(1 + \eta\,\ln 2 \sum_{j=2}^{N} \Big(\frac{j}{N}\Big)^{j-1}\Big) \;\le\; M\Big(1 + 1.65\,\eta\Big). \end{aligned} \tag{7.13}$$

One can check that the sum in the lower formula of (7.13) is maximal for $N = 7$ and bounded by 2.38. For a pth order method we obtain under the same assumptions

$$\|F_N(y) - f(y)\| \le cMh^p, \tag{7.14}$$

where c depends only on the method.

Theorem 7.6. *Let $f(y)$ be analytic in $B_{2R}(y_0)$, let the coefficients $d_j(y)$ of the method (7.3) be analytic in $B_R(y_0)$, and assume that (7.2) and (7.5) hold. If $h \le h_0/4$ with $h_0 = R/(e\eta M)$, then there exists $N = N(h)$ (namely N equal to the largest integer satisfying $hN \le h_0$) such that the difference between the numerical solution $y_1 = \Phi_h(y_0)$ and the exact solution $\widetilde{\varphi}_{N,t}(y_0)$ of the truncated modified equation (7.11) satisfies*

$$\|\Phi_h(y_0) - \widetilde{\varphi}_{N,h}(y_0)\| \;\le\; h\gamma M e^{-h_0/h},$$

where $\gamma = e(2 + 1.65\eta + \mu)$ depends only on the method (we have $5 \le \eta \le 5.18$ and $\gamma \le 31.4$ for the methods of Table 7.1).

The quotient $L = M/R$ is an upper bound of the first derivative $f'(y)$ and can be interpreted as a Lipschitz constant for $f(y)$. The condition $h \le h_0/4$ is therefore equivalent to $hL \le Const$, where $Const$ depends only on the method. Because of this condition, Theorem 7.6 requires unreasonably small step sizes for the numerical solution of stiff differential equations.

Proof of Theorem 7.6. We follow here the elegant proof of Benettin & Giorgilli (1994). It is based on the fact that $\Phi_h(y_0)$ (as a convergent series (7.3)) and $\widetilde{\varphi}_{N,h}(y_0)$ (as the solution of an analytic differential equation) are both analytic functions of h. Hence,

$$g(h) := \Phi_h(y_0) - \widetilde{\varphi}_{N,h}(y_0) \tag{7.15}$$

is analytic in a complex neighbourhood of $h = 0$. By definition of the functions $f_j(y)$ of the modified equation (1.1), the coefficients of the Taylor series for $\Phi_h(y_0)$ and $\widetilde{\varphi}_{N,h}(y_0)$ are the same up to the h^N term, but not further due to the truncation of the modified equation. Consequently, the function $g(h)$ contains the factor h^{N+1},

and the maximum principle for analytic functions, applied to $g(h)/h^{N+1}$, implies that

$$\|g(h)\| \le \left(\frac{h}{\varepsilon}\right)^{N+1} \max_{|z|\le\varepsilon} \|g(z)\| \qquad \text{for} \quad 0 \le h \le \varepsilon, \tag{7.16}$$

if $g(z)$ is analytic for $|z| \le \varepsilon$. We shall show that we can take $\varepsilon = eh_0/N$, and we compute an upper bound for $\|g(z)\|$ by estimating separately $\|\Phi_h(y_0) - y_0\|$ and $\|\widetilde{\varphi}_{N,h}(y_0) - y_0\|$.

The function $\Phi_z(y_0)$ is given by the series (7.3) which, due to the bounds of Theorem 7.2, converges certainly for $|z| \le R/(4\kappa M)$, and therefore also for $|z| \le \varepsilon$ (because $2\kappa \le \eta$ and $N \ge 4$, which is a consequence of $h_0/h \ge 4$). Hence, it is analytic in $|z| \le \varepsilon$. Moreover, we have from Theorem 7.2 that $\|\Phi_z(y_0) - y_0\| \le |z|M(1+\mu)$ for $|z| \le \varepsilon$.

Because of the bound (7.13) on $F_N(y)$, which is valid for $y \in B_{R/2}(y_0)$ and for $|h| \le \varepsilon$, we have $\|\widetilde{\varphi}_{N,z}(y_0) - y_0\| \le |z|M(1 + 1.65\eta)$ as long as the solution $\widetilde{\varphi}_{N,z}(y_0)$ stays in the ball $B_{R/2}(y_0)$. Because of $\varepsilon M(1+1.65\eta) \le R/2$, which is a consequence of the definition of ε, of $N \ge 4$, and of $(1 + 1.65\eta) \le 1.85\eta$ (because for consistent methods $\mu \ge 1$ holds and therefore also $\eta \ge 2/(2\ln 2 - 1) \ge 5$), this is the case for all $|z| \le \varepsilon$. In particular, the solution $\widetilde{\varphi}_{N,z}(y_0)$ is analytic in $|z| \le \varepsilon$.

Inserting $\varepsilon = eh_0/N$ and the bound on $\|g(z)\| \le \|\Phi_z(y_0) - y_0\| + \|\widetilde{\varphi}_{N,z}(y_0) - y_0\|$ into (7.16) yields (with $C = 2 + 1.65\eta + \mu$)

$$\|g(h)\| \le \varepsilon MC\left(\frac{h}{\varepsilon}\right)^{N+1} \le hMC\left(\frac{h}{\varepsilon}\right)^{N} = hMC\left(\frac{hN}{eh_0}\right)^{N} \le hMCe^{-N},$$

because $hN \le h_0$. The statement now follows from the fact that $N \le h_0/h < N+1$, so that $e^{-N} \le e \cdot e^{-h_0/h}$. □

A different approach to a rigorous backward error analysis is developed by Moan (2005). There, the modified differential equation contains an exponentially small time-dependent perturbation, but its flow reproduces the numerical solution without error.

IX.8 Long-Time Energy Conservation

> In particular, one easily explains in this way why symplectic algorithms give rise to a good energy conservation, with essentially no accumulation of errors in time. (G. Benettin & A. Giorgilli 1994)

As a first application of Theorem 7.6 we study the long-time energy conservation of symplectic numerical schemes applied to Hamiltonian systems $\dot{y} = J^{-1}\nabla H(y)$. It follows from Theorem 3.1 that the corresponding modified differential equation is also Hamiltonian. After truncation we thus get a modified Hamiltonian

$$\widetilde{H}(y) = H(y) + h^p H_{p+1}(y) + \ldots + h^{N-1}H_N(y), \tag{8.1}$$

which we assume to be defined on the same open set as the original Hamiltonian H; see Theorem 3.2 and Sect. IX.4. We also assume that the numerical method satisfies the analyticity bounds (7.5), so that Theorem 7.6 can be applied. The following result is given by Benettin & Giorgilli (1994).

Theorem 8.1. *Consider a Hamiltonian system with analytic $H : D \to \mathbb{R}$ (where $D \subset \mathbb{R}^{2d}$), and apply a symplectic numerical method $\Phi_h(y)$ with step size h. If the numerical solution stays in the compact set $K \subset D$, then there exist h_0 and $N = N(h)$ (as in Theorem 7.6) such that*

$$\begin{aligned} \widetilde{H}(y_n) &= \widetilde{H}(y_0) + \mathcal{O}(e^{-h_0/2h}) \\ H(y_n) &= H(y_0) + \mathcal{O}(h^p) \end{aligned}$$

over exponentially long time intervals $nh \le e^{h_0/2h}$.

Proof. We let $\widetilde{\varphi}_{N,t}(y_0)$ be the flow of the truncated modified equation. Since this differential equation is Hamiltonian with $\widetilde{H}$ of (8.1), $\widetilde{H}\big(\widetilde{\varphi}_{N,t}(y_0)\big) = \widetilde{H}(y_0)$ holds for all times t. From Theorem 7.6 we know that $\|y_{n+1} - \widetilde{\varphi}_{N,h}(y_n)\| \le h\gamma M e^{-h_0/h}$ and, by using a global h-independent Lipschitz constant for $\widetilde{H}$ (which exists by Theorem 7.5), we also get $\widetilde{H}(y_{n+1}) - \widetilde{H}(\widetilde{\varphi}_{N,h}(y_n)) = \mathcal{O}(he^{-h_0/h})$. From the identity

$$\widetilde{H}(y_n) - \widetilde{H}(y_0) = \sum_{j=1}^{n} \Big(\widetilde{H}(y_j) - \widetilde{H}(y_{j-1}) \Big) = \sum_{j=1}^{n} \Big(\widetilde{H}(y_j) - \widetilde{H}\big(\widetilde{\varphi}_{N,h}(y_{j-1})\big) \Big)$$

we thus get $\widetilde{H}(y_n) - \widetilde{H}(y_0) = \mathcal{O}(nhe^{-h_0/h})$, and the statement on the long-time conservation of $\widetilde{H}$ is an immediate consequence. The statement for the Hamiltonian H follows from (8.1), because $H_{p+1}(y) + hH_{p+2}(y) + \ldots + h^{N-p-1}H_N(y)$ is uniformly bounded on K independently of h and N. This follows from the proof of Lemma VI.2.7 and from the estimates of Theorem 7.5. □

Example 8.2. Let us check explicitly the assumptions of Theorem 8.1 for the pendulum problem $\dot{q} = p$, $\dot{p} = -\sin q$. The vector field $f(p,q) = (p, -\sin q)^T$ is also well-defined for complex p and q, and it is analytic everywhere on $\mathbb{C}^2$. We let K be a compact subset of $\{(p,q) \in \mathbb{R}^2 \,;\, |p| \le c\}$. As a consequence of $|\sin q| \le e^{|\mathrm{Im}\, q|}$, we get the bound

$$\|f(p,q)\| \le \sqrt{c^2 + 4R^2 + e^{2R}}$$

for $\|(p,q) - (p_0,q_0)\| \le 2R$ and $(p_0,q_0) \in K$. If we choose $c \le 2$, $R = 1$, and $M = 4$, the value h_0 of Theorem 7.6 is given by $h_0 = 1/4e\eta \approx 0.018$ for the methods of Table 7.1. For step sizes that are smaller than $h_0/20$, Theorem 8.1 guarantees that the numerical Hamiltonian is well conserved on intervals $[0,T]$ with $T \approx e^{10} \approx 2 \cdot 10^4$.

The numerical experiment of Fig. 8.1 shows that the estimates for h_0 are often too pessimistic. We have drawn 200 000 steps of the numerical solution of the

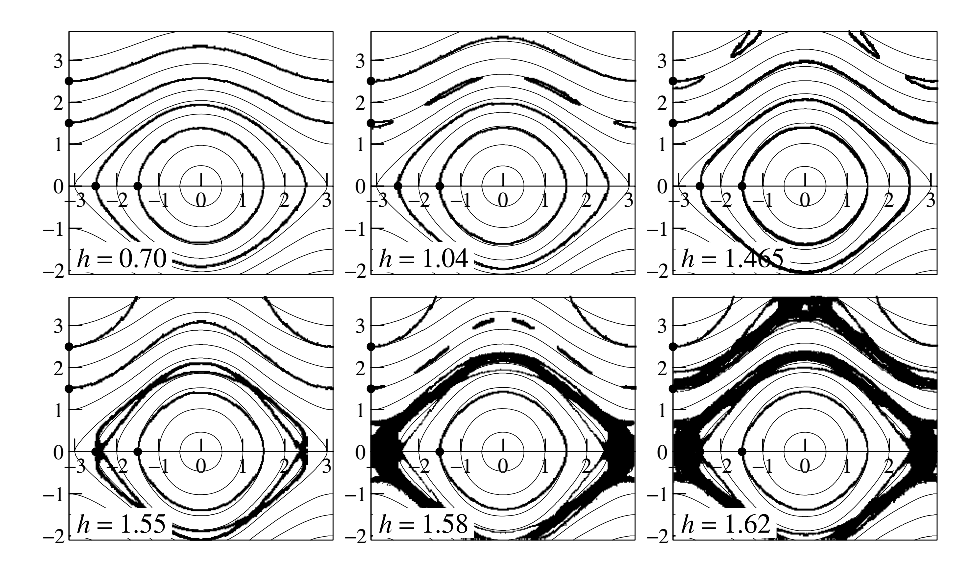


Fig. 8.1. Numerical solutions of the implicit midpoint rule with large step sizes

implicit midpoint rule for various step sizes h and for initial values $(p_0, q_0) = (0, -1.5)$, $(p_0, q_0) = (0, -2.5)$, $(p_0, q_0) = (1.5, -\pi)$, and $(p_0, q_0) = (2.5, -\pi)$. They are compared to the contour lines of the truncated modified Hamiltonian

$$\widetilde{H}(p, q) = \frac{p^2}{2} - \cos q + \frac{h^2}{48}\Big(\cos(2q) - 2p^2 \cos q\Big).$$

This shows that for step sizes as large as $h \leq 0.7$ the Hamiltonian $\widetilde{H}$ is extremely well conserved. Beyond this value, the dynamics of the numerical method soon turns into chaotic behaviour (see also Yoshida (1993) and Hairer, Nørsett & Wanner (1993), page 336).

Theorem 8.1 explains the near conservation of the Hamiltonian with the symplectic Euler method, the implicit midpoint rule and the Störmer–Verlet method as observed in the numerical experiments of Chap. I: in Fig. I.1.4 for the pendulum problem, in Fig. I.2.3 for the Kepler problem, and in Fig. I.4.1 for the frozen argon crystal.

The linear drift of the numerical Hamiltonian for non-symplectic methods can be explained by a computation similar to that of the proof of Theorem 8.1. From a Lipschitz condition of the Hamiltonian and from the standard local error estimate, we obtain $H(y_{n+1}) - H(\varphi_h(y_n)) = \mathcal{O}(h^{p+1})$. Since $H(\varphi_h(y_n)) = H(y_n)$, a summation of these terms leads to

$$H(y_n) - H(y_0) = \mathcal{O}(th^p) \qquad \text{for } t = nh. \tag{8.2}$$

This explains the linear growth in the error of the Hamiltonian observed in Fig. I.2.3 and in Fig. I.4.1 for the explicit Euler method.

IX.9 Modified Equation in Terms of Trees

By Theorem III.1.4 the numerical solution $y_1 = \Phi_h(y_0)$ of a Runge–Kutta method can be written as a B-series

$$\begin{aligned}\Phi_h(y) \;&=\; y \;+\; hf(y) + h^2\, a([\bullet])(f'f)(y)\\ &+\; h^3\Big(\frac{1}{2}\, a([\bullet,\bullet]) f''(f,f)(y) + a([[\bullet]]) f'f'f(y)\Big) + \ldots\;.\end{aligned} \tag{9.1}$$

For consistent methods, i.e., methods of order at least 1, we always have $a(\bullet) = 1$, so that the coefficient of h is equal to $f(y)$. In this section we exploit this special structure of $\Phi_h(y)$ in order to get practical formulas for the coefficient functions of the modified differential equation. Using (9.1) instead of (1.3), the equations (1.4) yield

$$\begin{aligned}f_2(y) \;&=\; \Big(a([\bullet]) - \frac{1}{2}\Big)\,(f'f)(y)\\ f_3(y) \;&=\; \frac{1}{2}\Big(a([\bullet,\bullet]) - a([\bullet]) + \frac{1}{6}\Big) f''(f,f)(y)\\ &+\;\; \Big(a([[\bullet]]) - a([\bullet]) + \frac{1}{3}\Big) f'f'f(y).\end{aligned} \tag{9.2}$$

Continuing this computation, one is quickly convinced of the general formula

$$f_j(y) \;=\; \sum_{|\tau|=j} \frac{b(\tau)}{\sigma(\tau)}\, F(\tau)(y), \tag{9.3}$$

so that the modified equation (1.1) becomes

$$\dot{\widetilde{y}} = \sum_{\tau\in T} \frac{h^{|\tau|-1}}{\sigma(\tau)}\, b(\tau)\, F(\tau)(\widetilde{y}\,) \tag{9.4}$$

with $b(\bullet) = 1$, $b([\bullet]) = a([\bullet]) - \frac{1}{2}$, etc. Since the coefficients $\sigma(\tau)$ are known from Definition III.1.7, all we have to do is to find suitable recursion formulas for the real coefficients $b(\tau)$.

IX.9.1 B-Series of the Modified Equation

Recurrence formulas for the coefficients $b(\tau)$ in (9.4) were first given by Hairer (1994) and by Calvo, Murua & Sanz-Serna (1994). We follow here the approach of Hairer (1999), which uses the Lie-derivative of B-series and thus simplifies the construction of the coefficients.

We make use of the notion of ordered trees introduced in Sect. III.1.3. For a given tree τ we define the set of all *splittings* as

$$SP(\tau) = \big\{\theta \in OST(\tau)\;;\; \tau\setminus\theta \;\text{ consists of only one element}\big\}. \tag{9.5}$$

Here, $OST(\tau) = OST\big(\omega(\tau)\big)$ is the set of ordered subtrees as defined in (III.1.33).

Lemma 9.1 (Lie-Derivative of B-series). *Let $b(\tau)$ (with $b(\emptyset) = 0$) and $c(\tau)$ be the coefficients of two B-series, and let $y(t)$ be a formal solution of the differential equation $h\dot{y}(t) = B\big(b, y(t)\big)$. The Lie derivative of the function $B(c, y)$ with respect to the vector field $B(b, y)$ is again a B-series*

$$h\,\frac{d}{dt}\,B\big(c, y(t)\big) = B\big(\partial_b c, y(t)\big).$$

Its coefficients are given by $\partial_b c(\emptyset) = 0$ and for $|\tau| \geq 1$ by

$$\partial_b c(\tau) = \sum_{\theta \in SP(\tau)} c(\theta)\, b(\tau \setminus \theta). \tag{9.6}$$

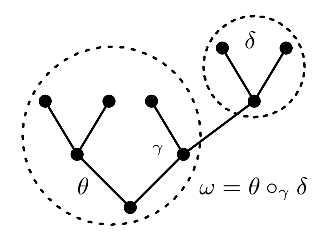


Fig. 9.1. Splitting of an ordered tree ω into a subtree θ and $\{\delta\} = \omega \setminus \theta$

Proof. For the proof of this lemma it is convenient to work with ordered trees $\omega \in OT$. Since $\nu(\tau)$ of (III.1.31) denotes the number of possible orderings of a tree $\tau \in T$, a sum $\sum_{\tau \in T} \cdot/\cdot$ becomes $\sum_{\omega \in OT} \nu(\omega)^{-1} \cdot /\cdot$.

For the computation of the Lie derivative of $B(c, y)$ we have to differentiate the elementary differential $F(\theta)\big(y(t)\big)$ with respect to t. Using Leibniz' rule, this yields $|\theta|$ terms, one for every vertex of θ. Then we insert the series $B\big(b, y(t)\big)$ for $h\dot{y}(t)$. This means that all the trees δ appearing in $B\big(b, y(t)\big)$ are attached with a new branch to the distinguished vertex. Written out as formulas, this gives

$$h\frac{d}{dt}B\big(c, y(t)\big) \;=\; \sum_{\theta \in OT \cup \{\emptyset\}} \frac{h^{|\theta|}\, c(\theta)}{\nu(\theta)\,\sigma(\theta)} \sum_{\gamma} \sum_{\delta \in OT} \frac{h^{|\delta|}\, b(\delta)}{\nu(\delta)\,\sigma(\delta)}\, F(\theta \circ_\gamma \delta)\big(y(t)\big),$$

where $\sum_\gamma$ is a sum over all vertices of θ, and $\theta \circ_\gamma \delta$ is the ordered tree obtained when attaching the root of δ with a new branch to γ (see Fig. 9.1). We choose one of the $n(\gamma)+1$ possibilities of doing this, where $n(\gamma)$ denotes the number of upwards leaving branches of θ at the vertex γ. We now collect the terms with equal ordered tree $\omega = \theta \circ_\gamma \delta$, and notice that $\nu(\theta)\sigma(\theta) = \kappa(\theta)$ with $\kappa(\theta)$ given by (III.1.32). This gives

$$h\frac{d}{dt}B\big(c, y(t)\big) \;=\; \sum_{\omega \in OT} h^{|\omega|} \Big(\sum_{\theta \circ_\gamma \delta = \omega} \frac{c(\theta)\, b(\delta)}{(n(\gamma)+1)\,\kappa(\theta)\,\kappa(\delta)} \Big) F(\omega)\big(y(t)\big),$$

where $\sum_{\theta \circ_\gamma \delta = \omega}$ is over all triplets (θ, γ, δ) such that $\theta \circ_\gamma \delta = \omega$. Because of $\kappa(\omega) = \kappa(\theta)\kappa(\delta)(n(\gamma)+1)$, we obtain

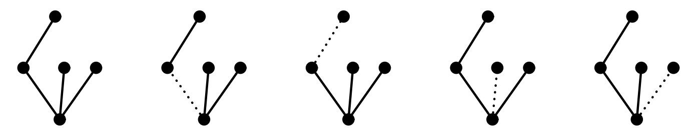


Fig. 9.2. Illustration of the formula (9.6) for an ordered tree with 5 vertices

$$
\begin{aligned}
h\frac{d}{dt}B\big(c,y(t)\big) &= \sum_{\omega\in OT}\frac{h^{|\omega|}}{\kappa(\omega)}\Big(\sum_{\theta\circ_\gamma\delta=\omega}c(\theta)\,b(\delta)\Big)F(\omega)\big(y(t)\big)\\
&= \sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\Big(\sum_{\theta\in SP(\tau)}c(\theta)\,b(\tau\setminus\theta)\Big)F(\tau)\big(y(t)\big),
\end{aligned}
$$

which proves the statement. □

Let us illustrate this proof and the formula (9.6) with an ordered tree having 5 vertices. All possible splittings $\omega = \theta\circ_\gamma\delta$ are given in Fig. 9.2. Notice that θ may be the empty tree $\emptyset$, and that always $|\delta|\geq 1$. We see that the tree ω is obtained in several ways: (i) differentiation of $F(\emptyset)(y) = y$ and adding $F(\omega)(y)$ as argument, (ii) differentiation of the factor corresponding to the root in $F(\theta)(y) = f''(f,f)(y)$ and adding $F([\bullet])(y) = (f'f)(y)$, (iii) differentiation of all f's in $F(\theta)(y) = f'''(f,f,f)(y)$ and adding $F(\bullet)(y) = f(y)$, and finally, (iv) differentiation of the factor for the root in $F(\theta)(y) = f''(f'f,f)(y)$ and adding $F(\bullet)(y) = f(y)$. This proves that

$$
\partial_b c([[\bullet],\bullet,\bullet]) = c(\emptyset)b([[\bullet],\bullet,\bullet]) + c([\bullet,\bullet])b([\bullet]) + c([\bullet,\bullet,\bullet])b(\bullet) + 2\,c([[\bullet],\bullet])b(\bullet).
$$

For the trees up to order 3 the formulas for $\partial_b c$ are:

$$
\begin{aligned}
\partial_b c(\bullet) &= c(\emptyset)\,b(\bullet)\\
\partial_b c([\bullet]) &= c(\emptyset)\,b([\bullet]) + c(\bullet)\,b(\bullet)\\
\partial_b c([\bullet,\bullet]) &= c(\emptyset)\,b([\bullet,\bullet]) + 2\,c([\bullet])\,b(\bullet)\\
\partial_b c([[\bullet]]) &= c(\emptyset)\,b([[\bullet]]) + c(\bullet)\,b([\bullet]) + c([\bullet])\,b(\bullet).
\end{aligned}
$$

The above lemma permits us to get recursion formulas for the coefficients $b(\tau)$ of the modified differential equation (9.4).

Theorem 9.2. *If the method $\Phi_h(y)$ is given by (9.1), the functions $f_j(y)$ of the modified differential equation (1.1) satisfy (9.3), where the real coefficients $b(\tau)$ are recursively defined by $b(\emptyset)=0$, $b(\bullet)=1$ and*

$$
b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|}\frac{1}{j!}\,\partial_b^{j-1}b(\tau). \tag{9.7}
$$

Here, ∂_b^{j-1} is the $(j-1)$-th iterate of the Lie-derivative ∂_b defined in Lemma 9.1.

Proof. The right-hand side of the modified equation (9.4) is the B-series $B\big(b, \widetilde{y}(t)\big)$ divided by h. It therefore follows from an iterative application of Lemma 9.1 that

$$h^j \, \widetilde{y}^{\,(j)}(t) = B\big(\partial_b^{j-1} b, \widetilde{y}(t)\big),$$

so that by Taylor series expansion $\widetilde{y}(t+h) = y + B\big(\sum_{j\ge 1} \frac{1}{j!}\, \partial_b^{j-1} b, y\big)$, where $y := \widetilde{y}(t)$. Since we have to determine the coefficients $b(\tau)$ in such a way that $\widetilde{y}(t+h) = \Phi_h(y) = B(a,y)$, a comparison of the two B-series gives $\sum_{j\ge 1} \frac{1}{j!}\, \partial_b^{j-1} b(\tau) = a(\tau)$. This proves the statement, because $\partial_b^0 b(\tau) = b(\tau)$ for $\tau \in T$, and $\partial_b^{j-1} b(\tau) = 0$ for $j > |\tau|$ (as a consequence of $b(\emptyset) = 0$). □

We present in Table 9.1 the formula (9.7) for trees up to order 3.

Table 9.1. Examples of formula (9.7)

$\tau = \bullet$	$b(\bullet) = a(\bullet)$
$\tau = [\bullet]$	$b([\bullet]) = a([\bullet]) - \frac{1}{2}\, b(\bullet)^2$
$\tau = [\bullet,\bullet]$	$b([\bullet,\bullet]) = a([\bullet,\bullet]) - b([\bullet])b(\bullet) - \frac{1}{3}\, b(\bullet)^3$
$\tau = [[\bullet]]$	$b([[\bullet]]) = a([[\bullet]]) - b([\bullet])b(\bullet) - \frac{1}{6}\, b(\bullet)^3$

We next consider the case when a symplectic method is applied to a Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$. It follows from Theorem 3.1 that the modified equation is again Hamiltonian. What does this imply for the coefficients of (9.4)?

Theorem 9.3. *Suppose that for all Hamiltonians $H(y)$ the modified vector field (9.4), truncated after an arbitrary power of h, is (locally) Hamiltonian. Then,*

$$b(u \circ v) + b(v \circ u) \;=\; 0 \qquad \textit{for all} \quad u,\, v \in T. \tag{9.8}$$

Proof. Let $\widetilde{\varphi}_{N,t}(y_0)$ be the flow of the modified differential equation (9.4), truncated after the h^{N-1} terms. It is symplectic for all t, and in particular for $t = h$. As a consequence of the proof of Theorem 9.2 we obtain that $\widetilde{\varphi}_{N,h}(y_0)$ is a symplectic B-series $B\big(a_N, y_0\big)$. The coefficients $a_N(\tau)$ are given by (9.7), where $b(\tau)$ is replaced with 0 for $|\tau| > N$. For $u, v \in T$ with $|u| + |v| = N$ we therefore have

$$b(u \circ v) = a_N(u \circ v) - a_{N-1}(u \circ v).$$

Since $a_N(\tau) = a_{N-1}(\tau)$ for $|\tau| < N$, formula (9.8) is an immediate consequence of Theorem VI.7.6. □

Remark 9.4. Let $G = \{a : T \to \mathbb{R} \mid a(\emptyset) = 1\}$ be the Butcher group (see Sect. III.1.5), and consider the mapping $S : G \to \mathbb{R}$ defined by

$$S(a) \;=\; a(u \circ v) + a(v \circ u) - a(u) \cdot a(v).$$

If we denote by $e \in G$ the element corresponding to the identity (i.e., $e(\emptyset) = 1$ and $e(\tau) = 0$ for $|\tau| \geq 1$), we have for its derivative

$$S'(e)b \,=\, b(u \circ v) + b(v \circ u).$$

Hence, coefficient mappings $b(\tau)$ satisfying (9.8) lie in the tangent space at $e(\tau)$ of the symplectic subgroup of G (i.e., $a \in G$ satisfying (VI.7.4)). This is in complete analogy to the fact that Hamiltonian vector fields can be considered as elements of the tangent space at the identity of the group of symplectic diffeomorphisms (see also Exercises 15 and 16).

IX.9.2 Elementary Hamiltonians

If the modified differential equation (9.4) is Hamiltonian, can we find explicit formulas for $\widetilde{H}(y)$? Let us start with an easy example, the implicit midpoint rule. Written as a B-series (9.1), its coefficients are $a(\tau) = 2^{1-|\tau|}$ (cf. Exercise 8) so that the first coefficient functions (9.2) of the modified equation satisfy $f_2(y) = 0$ and

$$f_3(y) = \frac{1}{24}\Big(2(f'f'f)(y) - f''(f,f)(y)\Big). \tag{9.9}$$

Since $f(y) = J^{-1}\nabla H(y)$, differentiation of

$$H_3(y) = -\frac{1}{24}\, H''(y)\Big(J^{-1}\nabla H(y), J^{-1}\nabla H(y)\Big) \tag{9.10}$$

shows that $f_3(y) = J^{-1}\nabla H_3(y)$, and we have found an explicit expression of the Hamiltonian corresponding to the vector field $f_3(y)$. It is recommended to compute also $f_5(y)$ and to try to find $H_5(y)$ such that $f_5(y) = J^{-1}\nabla H_5(y)$. Such computations lead to expressions that have been introduced in a different context by Sanz-Serna & Abia (1991). They call them *canonical elementary differentials*.

Definition 9.5 (Elementary Hamiltonians). For a given smooth function $H : D \to \mathbb{R}$ (with open $D \subset \mathbb{R}^{2d}$) and for $\tau \in T$ we define the *elementary Hamiltonian* $H(\tau) : D \to \mathbb{R}$ by

$$H(\bullet)(y) = H(y), \qquad H(\tau)(y) = H^{(m)}(y)\big(F(\tau_1)(y), \ldots, F(\tau_m)(y)\big) \tag{9.11}$$

for $\tau = [\tau_1, \ldots, \tau_m]$. Here, $F(\tau_i)(y)$ are elementary differentials corresponding to $f(y) = J^{-1}\nabla H(y)$.

The expression in (9.10) is nothing else than the elementary Hamiltonian corresponding to the tree $\vee$. Our aim is to prove that, for symplectic methods applied to Hamiltonian systems, the coefficient functions (9.3) of the modified differential equation satisfy $f_j(y) = J^{-1}\nabla H_j(y)$, where $H_j(y)$ is a linear combination of elementary Hamiltonians.

Lemma 9.6. *Elementary Hamiltonians satisfy*

$$H(u \circ v)(y) + H(v \circ u)(y) = 0 \qquad \textit{for all} \;\; u, v \in T. \tag{9.12}$$

In particular, we have $H(u \circ u)(y) = 0$ *for all* $u \in T$.

Proof. This follows immediately from the fact that for $u = [u_1, \ldots, u_m] \in T$ and for $v \in T$ we have $H(u \circ v) = H^{(m+1)}\big(F(u_1), \ldots, F(u_m), F(v)\big) = F(v)^T (\nabla H)^{(m)}\big(F(u_1), \ldots, F(u_m)\big) = F(v)^T J\, F(u)$, and from the skew-symmetry of J. □

The trees $u \circ v$ and $v \circ u$ have the same graph and differ only in the position of the root. The relation (9.12) thus motivates the consideration of the (smallest) equivalence relation on T satisfying

$$u \circ v \sim v \circ u. \tag{9.13}$$

We want to select from each equivalence class, not containing a tree of the form $u \circ u$, exactly one element. This can be done as follows (cf. Chartier, Faou & Murua 2005): we choose a total ordering on the set T that respects the number of vertices, i.e., $u < v$ whenever $|u| < |v|$, and we define

$$T^* \;=\; \{\bullet\} \cup \{\tau \,|\, \tau \text{ cannot be written as } \tau = u \circ v \text{ with } u \leq v\} \tag{9.14}$$

$$= \Big\{ \bullet \quad \ldots \Big\}$$

(for the second line we assume $[\bullet, \bullet] < [[\bullet]]$). Every tree $\tau \in T$ is either equivalent to some $u \circ u$ or to a tree in T^*. This is a consequence of the fact that as long as $\tau = u \circ v$ with $u < v$, it can be changed to $v \circ u$ (what happens only a finite number of times). Moreover, two trees of T^* can never be equivalent.

Lemma 9.7. *For a tree* $\tau \in T^*$ *we have*

$$J^{-1} \nabla H(\tau)(y) = \sigma(\tau) \sum_{\theta \sim \tau} \frac{(-1)^{\kappa(\tau,\theta)}}{\sigma(\theta)}\, F(\theta)(y), \tag{9.15}$$

where $\kappa(\tau, \theta)$ *is the number of root changes that are necessary to obtain* θ *from* τ.

Proof. We compute $J^{-1}\nabla H(\tau)(y)$. The expression $H(\tau)(y)$ consists of $|\tau|$ factors corresponding to the vertices of τ, each of which has to be differentiated by Leibniz' rule. Differentiation of $H^{(m)}(y)$ (cf. Definition 9.5) and pre-multiplication by the matrix J^{-1} yields $F(\tau)(y)$. Before differentiating the other factors, we bring the corresponding vertex down to the root. In view of Lemma 9.6 this only multiplies $H(\tau)(y)$ by $(-1)^{\kappa(\tau,\theta)}$, and shows that a differentiation of the corresponding factor yields $F(\theta)(y)$. Since $\tau \in T^*$, the number of possibilities to obtain θ from τ by exchanging roots is equal to $\sigma(\tau)/\sigma(\theta)$. This factor has to be included. □

IX.9.3 Modified Hamiltonian

We are now in the position to give an explicit formula for the Hamiltonian of the modified differential equation provided that the numerical method can be written as a B-series. An extension to partitioned methods will be given in Sect. IX.10.

Theorem 9.8. *Consider a numerical method that can be written as a B-series (9.1), and that is symplectic for every Hamiltonian system* $\dot{y} = J^{-1}\nabla H(y)$. *Its modified differential equation is then Hamiltonian with*

$$\widetilde{H}(y) \;=\; H_1(y) + h\,H_2(y) + h^2\,H_3(y) + \ldots,$$

where

$$H_j(y) = \sum_{\tau\in T^*,\,|\tau|=j} \frac{b(\tau)}{\sigma(\tau)}\,H(\tau)(y), \tag{9.16}$$

and the coefficients $b(\tau)$ *are those of Theorem 9.2. Notice that the sum in (9.16) is only over trees in* T^* *as defined in (9.14).*

Proof. We apply the method (9.1) to the Hamiltonian system, so that by Theorem 3.1 the modified differential equation is (locally) Hamiltonian. It therefore follows from Theorem 9.3 that the coefficients $b(\tau)$ of (9.4) satisfy (9.8). This relation implies $b(\theta) = (-1)^{\kappa(\tau,\theta)} b(\tau)$ whenever $\theta \sim \tau$. Inserted into (9.3), an application of Lemma 9.7 proves the statement. □

Remark 9.9. This theorem gives an explicit formula for the modified Hamiltonian (for methods expressed as B-series). Since the elementary Hamiltonians $H(\tau)(y)$ depend only on derivatives of $H(y)$, this modified Hamiltonian is *globally* defined. For Runge–Kutta methods this provides an alternative approach to the statement of Theorem 3.2.

For the sake of completeness we give in the following theorem a characterization of Hamiltonian vector fields of the form (9.4).

Theorem 9.10. *The differential equation* $h\dot{y} = B(b, y)$ *with* $b(\emptyset) = 0$ *is Hamiltonian for all vector fields* $f(y) = J^{-1}\nabla H(y)$, *if and only if*

$$b(u \circ v) + b(v \circ u) = 0 \qquad \textit{for all} \quad u, v \in T. \tag{9.17}$$

Proof. The "only if" part follows from Theorem 9.3. The "if" part is a consequence of the proof of Theorem 9.8. □

IX.9.4 First Integrals Close to the Hamiltonian

We have seen in Sect. IX.9.3 that for symplectic methods the modified differential equation (9.4) based on $f(y) = J^{-1}\nabla H(y)$ is Hamiltonian with a function of the form

$$H(c,y)=\sum_{\tau\in T^*}\frac{h^{|\tau|-1}}{\sigma(\tau)}\,c(\tau)\,H(\tau)(y) \tag{9.18}$$

and coefficients $c(\tau)=b(\tau)$. In this section we study whether for non-symplectic methods a function of the form (9.18) can be a first integral of (9.4). This question has been addressed by Faou, Hairer & Pham (2004), and we closely follow their presentation.

Lemma 9.11. *Let $y(t)$ be a solution of the differential equation (9.4) which can be written as $h\dot y(t)=B(b,y(t))$. We then have*

$$\frac{d}{dt}H\big(c,y(t)\big)=H\big(\delta_b c,y(t)\big)$$

where $\delta_c b(\bullet)=0$ *and, for $\tau\in T^*$ with $|\tau|>1$,*

$$\delta_b c(\tau)=\sum_{\theta\sim\tau}(-1)^{\kappa(\tau,\theta)}\,\frac{\sigma(\tau)}{\sigma(\theta)}\sum_{\omega\in T^*\cap SP(\theta)}c(\omega)\,b(\theta\setminus\omega). \tag{9.19}$$

The first sum is over all trees θ that are equivalent to τ (see (9.13)), and the second sum is over all splittings of θ as in Lemma 9.1 (see Table 9.2).

Proof. The proof is nearly the same as that of Lemma 9.1. The first sum in (9.19) appears, because $H(\theta)(y)=H(\tau)(y)$ for $\theta\sim\tau$ and because the sum in (9.18) is only over trees in T^*. □

Table 9.2. Formulas for $\delta_b c(\tau)$ for trees $\tau\in T^*$ up to order 6

$$\begin{aligned}
\delta_b c([\bullet,\bullet]) &= -2\,c(\bullet)b([\bullet])\\
\delta_b c([\bullet,\bullet,\bullet]) &= 3\,c([\bullet,\bullet])b(\bullet)-3\,c(\bullet)b([\bullet,\bullet])\\
\delta_b c([\bullet,\bullet,\bullet,\bullet]) &= 4\,c([\bullet,\bullet,\bullet])b(\bullet)-4\,c(\bullet)b([\bullet,\bullet,\bullet])\\
\delta_b c([[\bullet],\bullet,\bullet]) &= c([\bullet,\bullet,\bullet])b(\bullet)+c([\bullet,\bullet])b([\bullet])+c(\bullet)b([[\bullet,\bullet]])-2\,c(\bullet)b([[\bullet],\bullet])\\
\delta_b c([[\bullet],[\bullet]]) &= 2\,c(\bullet)b([[[\bullet]]])-2\,c([\bullet,\bullet])b([\bullet])\\
\delta_b c([\bullet,\bullet,\bullet,\bullet,\bullet]) &= 5\,c([\bullet,\bullet,\bullet,\bullet])b(\bullet)-5\,c(\bullet)b([\bullet,\bullet,\bullet,\bullet])\\
\delta_b c([[\bullet],\bullet,\bullet,\bullet]) &= 3\,c([[\bullet],\bullet,\bullet])b(\bullet)+c([\bullet,\bullet,\bullet,\bullet])b(\bullet)+c([\bullet,\bullet,\bullet])b([\bullet])\\
&\quad -3\,c(\bullet)b([[\bullet],\bullet,\bullet])+c(\bullet)b([[\bullet,\bullet,\bullet]])\\
\delta_b c([[\bullet],[\bullet],\bullet]) &= 2\,c([[\bullet],\bullet,\bullet])b(\bullet)+c([[\bullet],[\bullet]])b(\bullet)-c(\bullet)b([[\bullet],[\bullet]])+2\,c(\bullet)b([[[\bullet],\bullet]])\\
\delta_b c([[\bullet],[\bullet,\bullet]]) &= 2c([[\bullet],[\bullet]])b(\bullet)-c([[\bullet],\bullet,\bullet])b(\bullet)-c([\bullet,\bullet,\bullet])b([\bullet])-c([\bullet,\bullet])b([\bullet,\bullet])\\
&\quad -c([\bullet,\bullet])b([[\bullet]])+2\,c(\bullet)b([[[\bullet]],\bullet])+c(\bullet)b([[[\bullet,\bullet]]])
\end{aligned}$$

Corollary 9.12. *The function $H(c,y)$ of (9.18) is a first integral of the differential equation (9.4) for every $H(y)$ if and only if*

$$\delta_b c(\tau) = 0 \quad \textit{for all } \tau \in T^*. \tag{9.20}$$

Proof. The sufficiency follows from Lemma 9.11 and the necessity is a consequence of the independence of the elementary Hamiltonians. To prove their independence we have to show that the series (9.18) vanishes for all smooth $H(y)$ only if $c(\tau) = 0$ for all $\tau \in T^*$. With the techniques of the proof of Theorem VI.7.4 one can show that for every tree $\tau \in T^*$ there exists a polynomial Hamiltonian such that the first component of $F(\tau)(0)$ vanishes for all trees except for τ. Differentiating (9.18) and employing Lemma 9.7 proves that $c(\tau) = 0$. □

Solving the System (9.20). We consider a consistent method, i.e., $b(\bullet) = 1$, and we search for a first integral $H(c,y)$ close to the Hamiltonian, i.e., $c(\bullet) = 1$.

$|\tau| = 3$: The condition (9.20) for $\tau = [\bullet,\bullet]$ implies $b([\bullet]) = 0$, which means that the method has to be of order two.

$|\tau| = 4$: There is only one tree in T^* with four vertices. The corresponding condition can be satisfied by putting $c([\bullet,\bullet]) = b([\bullet,\bullet])$.

$|\tau| = 5$: The third condition yields $b([[[\bullet]]]) = 0$. Letting $c([\bullet,\bullet,\bullet])$ be such that one of the other two conditions holds, we still have to satisfy

$$b([\bullet,\bullet,\bullet]) + b([[\bullet,\bullet]]) - 2\,b([[\bullet],\bullet]) = 0. \tag{9.21}$$

This condition is satisfied for symplectic methods, for which $b(u \circ v) + b(v \circ u) = 0$, and also for symmetric methods, for which $b(\tau) = 0$ for trees with an even order.

$|\tau| = 6$: There are four conditions for three $c(\tau)$ coefficients. Assuming (9.20) for trees with less than five vertices, these four conditions admit a solution if and only if

$$\begin{aligned} 5\,b([\bullet,\bullet,\bullet,\bullet]) + 5\,b([[\bullet,\bullet,\bullet]]) + 6\,b([[\bullet],[\bullet]]) + 6\,b([[[\bullet]],\bullet]) - 12\,b([[[\bullet],\bullet]]) + 3\,b([[[\bullet,\bullet]]]) \\ -15\,b([[\bullet],\bullet,\bullet]) - 3\,b([\bullet,\bullet])\Big(b([\bullet,\bullet]) + b([[\bullet]])\Big) = 0. \end{aligned} \tag{9.22}$$

This relation is obviously satisfied by every symplectic method. However, as we shall see soon, there are symmetric methods that do not satisfy (9.22).

For various symmetric methods of order 4 (i.e., $b(\tau) = 0$ for $1 < |\tau| < 5$) we compute the coefficients $b(\tau)$ of the leading perturbation term in (9.4) and also the expression (9.22), see Table 9.3. None of the considered methods is symplectic.

Surprisingly, the 3-stage collocation method Lobatto IIIA (see Table II.1.2 for the coefficients) satisfies the condition (9.22). This implies for every Hamiltonian system (reversible or not reversible) that the dominating error term in the numerical Hamiltonian does not have any drift.

The 3-stage Lobatto IIIB method (see Table II.1.4) does not satisfy the condition (9.22). We therefore expect a drift in the numerical Hamiltonian.

Table 9.3. Coefficients $b(\tau)$ and expression (9.22) for methods of order 4

method								(9.22)
Lobatto IIIA	$\frac{1}{120}$	$\frac{1}{240}$	$\frac{1}{480}$	$-\frac{1}{120}$	$-\frac{1}{240}$	$\frac{1}{720}$	$-\frac{1}{360}$	0
Lobatto IIIB	$\frac{1}{120}$	$-\frac{1}{360}$	$-\frac{1}{720}$	$-\frac{1}{120}$	$\frac{1}{360}$	$\frac{1}{720}$	$\frac{1}{240}$	$\frac{1}{48}$

Lemma 9.13. *For given $b(\tau), \tau \in T$ satisfying $b(\emptyset) = 0$, $b(\bullet) = 1$, and for fixed $c(\bullet)$, the linear system (9.20) for $c(\tau), \tau \in T^*$ has at most one solution.*

Proof. We prove by induction on $\tau \in T^*$ that $c(\tau)$ is uniquely determined by (9.20). For this we assume that the ordering on T is such that, within trees of the same order, it is increasing when the numer of vertices connected to the root decreases, cf. (9.14).

Let $\tau = [\tau_1, \ldots, \tau_m, \bullet, \ldots, \bullet] \in T^* \setminus \{\bullet\}$ with $|\tau_j| > 1$, and denote by k the number of $\bullet$'s in this representation. Since the tree $\tau \circ \bullet$ is again in the set T^*, condition (9.20) yields

$$0 = \delta_b c(\tau \circ \bullet) = (k+1)c(\tau)b(\bullet) - (k+1)c(\bullet)b(\tau) + \ldots \, . \tag{9.23}$$

For $m = 0$, no further terms are present and $c(\tau)$ is uniquely determined by this relation. For $m > 0$, the three dots in (9.23) represent a linear combination of $c(\mu)b(\nu)$ with $|\mu| < |\tau|$ (which, by the induction hypothesis, are already known) and of $c(\sigma)b(\bullet)$, where $\sigma \in T^*$ is the representant in T^* of the equivalence class for τ'. We use the notation τ' for some tree which is obtained from τ by removing one of the end vertices of τ_j and by adding it to the root of τ.

In general we will have $\tau' \in T^*$ (so that $\sigma = \tau'$), and in this case its number of end vertices connected to the root is larger than that for τ. Hence, $\sigma < \tau$, and the coefficient $c(\sigma)$ is known by the induction hypothesis.

If $\tau' \notin T^*$, what is only possible if $\tau = u \circ v$ with $|u| = |v|$ and $u > v$, we have $\tau' = u' \circ v$ and $u' < v$ (notice that $u' = v$ is not permitted for trees in T^*). In this case we have $\sigma = v \circ u' \in T^*$. Consequently, $c(\tau) = c(u \circ v)$ is expressed in terms of $c(v \circ u')$ and known quantities. Applying the same reasoning to $v \circ u'$ and observing that because of $u > v$ the tree v has at least as many end vertices connected to the root as the tree u, we see that $c(v \circ u')$ is expressed in terms of already determined quantities. □

The expression (9.20) is bilinear in b and c. Assuming that $h\dot{y} = B(b, y)$ is Hamiltonian, the mapping b has the same degree of freedom as c. It is therefore not astonishing to have the following dual variant of Lemma 9.13.

Lemma 9.14. *Let $c(\tau), \tau \in T^*$ be given and assume $c(\bullet) = 1$ and $b(\emptyset) = 0$. Then, for fixed $b(\bullet)$, the linear system (9.20) for $b(\tau), \tau \in T$ has at most one solution satisfying $b(u \circ v) + b(v \circ u) = 0$ for all $u, v \in T$.*

Proof. By assumption on b, the coefficients $b(\tau)$, $\tau \in T \setminus T^*$ are uniquely determined by those for $\tau \in T^*$. The statement is thus obtained in the same way as that for Lemma 9.13 with the only difference that expressions $c(\bullet)b(\sigma)$ and not $c(\sigma)b(\bullet)$ have to be studied. □

Theorem 9.15 (Chartier, Faou & Murua 2005). *The only symplectic method (as B-series) that conserves the Hamiltonian for arbitrary $H(y)$ is the exact flow of the differential equation.*

Proof. If the method conserves exactly the Hamiltonian, we have (9.20) with $c(\bullet) = 1$ and $c(\tau) = 0$ for all other trees in T^*. By the uniqueness statement of Lemma 9.14 and the symplecticity of the method (Theorem 9.10), we obtain $b(\tau) = 0$ for $|\tau| > 1$. Consequently, no perturbation is permitted in the modified differential equation of the method. □

A closely related result is given in Ge & Marsden (1988). There, general symplectic methods are considered (not necessarily B-series methods) but a weaker result is obtained (in fact, they assume that the system does not have other conserved quantities than $H(y)$, and it is shown that the numerical flow coincides with the exact flow up to a reparametrization of time).

IX.9.5 Energy Conservation: Examples and Counter-Examples

It is generally believed that symmetric methods applied to reversible Hamiltonian systems (reversible in the sense that $H(-p,q) = H(p,q)$) have the same long-time behaviour as symplectic methods. This is true in many situations of practical interest, and we shall prove this rigorously in Sect. XI.3 for integrable reversible systems. There are, however, interesting counter-examples to this general belief. They are taken from Faou, Hairer & Pham (2004).

Example 9.16. Our first example is a modification of the pendulum equation

$$H(p,q) = \frac{1}{2}\,p^2 - \cos q + \frac{1}{5}\,\sin(2q), \tag{9.24}$$

where the additional term $\sin(2q)$ destroys the symmetry in q. The Hamiltonian still satisfies $H(-p,q) = H(p,q)$. We consider initial values $p(0) = 2.5$, $q(0) = 0$ with sufficiently large initial velocity, such that $p(t)$ stays positive for all times and the symmetry $p \leftrightarrow -p$ does not affect the numerical solution. The angle $q(t)$ increases without limit, but the potential is 2π-periodic so that the solution stays on a closed curve of the cylinder $\mathbb{R} \times S^1$.

We apply the 3-stage Lobatto IIIA and IIIB methods to this problem. Figure 9.3 shows the error in the Hamiltonian along the numerical solutions. There is a visible energy drift of size $\mathcal{O}(th^4)$ for the Lobatto IIIB method and no drift can be seen on this scale for the Lobatto IIIA method. To get more insight into its long-time behaviour, we apply the method with the same step size to a much longer time interval, and we plot the error in $H(p_n,q_n)+h^4H_5(p_n,q_n)$, where the first perturbation term is computed from (9.18) and the linear system (9.20) as

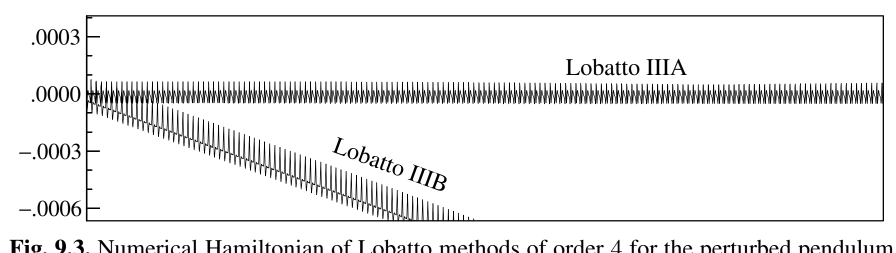


Fig. 9.3. Numerical Hamiltonian of Lobatto methods of order 4 for the perturbed pendulum (9.24); step size $h = 0.2$, integration interval $[0, 500]$

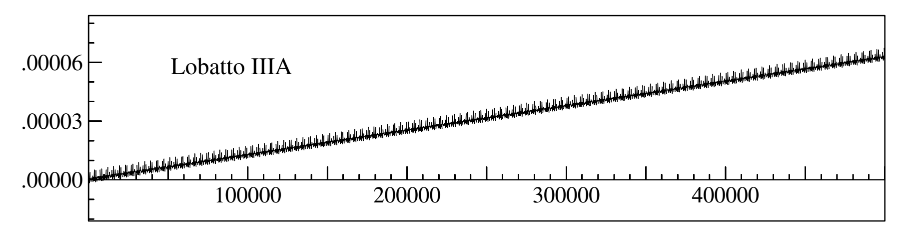


Fig. 9.4. Error in $H(p,q) + h^4 H_5(p,q)$ along the numerical solution of the 3-stage Lobatto IIIA method for the perturbed pendulum (9.24); step size $h = 0.2$, integration interval $[0, 500\,000]$

$$H_5(p,q) = \frac{1}{960}\Big(3\,U^{(4)}(q)p^4 - 2U^{(3)}U'(q)p^2 - \big(U''(q)p\big)^2 + U''(q)\big(U'(q)\big)^2\Big)$$

with the potential $U(q) = -\cos q + 0.2\sin(2q)$ (see Fig. 9.4). Repeating the same experiment with halved step size shows that there are oscillations with amplitude $\mathcal{O}(h^6)$ and a drift with slope $\mathcal{O}(h^8)$. Consequently, the error in the Hamiltonian for the Lobatto IIIA method behaves on this problem like $\mathcal{O}(h^4 + th^8)$.

Without the term $\sin(2q)$ in (9.24) all symmetric one-step methods nearly conserve the Hamiltonian.

Example 9.17. For polynomial Hamiltonians $H(y)$ of degree at most four, the elementary Hamiltonian corresponding to the tree [tree symbol] vanishes identically. Therefore, the condition (9.20) need not be considered for this tree, and the remaining three conditions can always be satisfied by the three $c(\tau)$ coefficients. This implies that, for example for the Hénon–Heiles problem

$$H(p_1,p_2,q_1,q_2) = \frac{1}{2}\big(p_1^2+p_2^2\big) + \frac{1}{2}\big(q_1^2+q_2^2\big) + q_1q_2^2 - \frac{1}{3}\,q_1^3, \tag{9.25}$$

the leading error term in the numerical Hamiltonian remains bounded by all methods of order four. Numerical experiments indicate that in this case also higher order error terms are bounded by symmetric methods such as Lobatto IIIA and IIIB, even if the initial values are chosen so that the solution is chaotic.

Example 9.18. A concrete mechanical system with two degrees of freedom is described by the Hamiltonian

$$H(p,q) = \frac{1}{2}\,p^T p + \frac{\omega^2}{2}\big(\|q\| - 1\big)^2 + q_2 - \frac{1}{\|q-a\|}. \tag{9.26}$$

It is a model of a planar spring pendulum with exterior forces. The spring has a harmonic potential with frequency ω (Hooke's law). The exterior forces are gravitation and attraction to a mass point situated at a, which has to be chosen so that no symmetry in the q-variables is present.

The numerical experiments, reported by Faou, Hairer & Pham (2004), use $\omega = 2$, $a = (-3,-5)^T$, and initial values for the position $q(0) = (0,1)^T$ (upright position), and for the velocity $p(0) = (-1,-0.5)^T$. The pendulum thus turns around the fixed end of the spring which is at the origin.

As for the problem of Example 9.16 one clearly observes a drift for the 3-stage Lobatto IIIB method, and the error in the Hamiltonian behaves like $\mathcal{O}(th^4)$. As predicted by the theory of the preceding section, the dominant error term for the 3-stage Lobatto IIIA method is bounded. There is, however, a drift already in the next term so that the error in the Hamiltonian behaves for this method as $\mathcal{O}(h^4 + th^6)$.

Removing one of the exterior forces (gravitation or attraction to a), the error in the Hamiltonian remains bounded of size $\mathcal{O}(h^4)$ without any drift (even not in higher order terms) for both Lobatto methods.

IX.10 Extension to Partitioned Systems

All results of Sect. IX.9 can be extended to partitioned methods whose discrete flow can be written as a P-series. This includes important geometric integrators such as the symplectic Euler method and the Störmer–Verlet scheme. Interestingly, many of the results have been originally presented and proved for this more general case (see Hairer (1994)).

IX.10.1 P-Series of the Modified Equation

We consider the partitioned system

$$\dot p = f(p,q), \qquad \dot q = g(p,q), \tag{10.1}$$

where, in view of an application to Hamiltonian systems, we use (p,q) instead of (y,z) for the variables. By Theorem III.2.4 all consistent partitioned Runge–Kutta methods can be written as P-series (cf. Definition III.2.1)

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + h \begin{pmatrix} f \\ g \end{pmatrix}_0 + h^2 \begin{pmatrix} a(\overset{\bullet}{\underset{\bullet}{|}})(f_p f) + a(\overset{\circ}{\underset{\bullet}{|}})(f_q g) \\ a(\overset{\bullet}{\underset{\circ}{|}})(g_p f) + a(\overset{\circ}{\underset{\circ}{|}})(g_q g) \end{pmatrix}_0 + \ldots, \tag{10.2}$$

where the subscript 0 indicates an evaluation at the initial value (p_0, q_0). The first perturbation term of the modified equation (1.1) can therefore be written as

$$\begin{pmatrix} f_2(p,q) \\ g_2(p,q) \end{pmatrix} = \begin{pmatrix} \big(a(\substack{\bullet\\ \bullet}) - \frac{1}{2}\big)(f_p f)(p,q) + \big(a(\substack{\circ\\ \bullet}) - \frac{1}{2}\big)(f_q g)(p,q) \\ \big(a(\substack{\bullet\\ \circ}) - \frac{1}{2}\big)(g_p f)(p,q) + \big(a(\substack{\circ\\ \circ}) - \frac{1}{2}\big)(g_q g)(p,q) \end{pmatrix}$$

and, in general, one finds

$$\begin{pmatrix} f_j(p,q) \\ g_j(p,q) \end{pmatrix} = \begin{pmatrix} \sum_{\tau\in TP_p, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)}\, F(\tau)(p,q) \\ \sum_{\tau\in TP_q, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)}\, F(\tau)(p,q) \end{pmatrix}. \tag{10.3}$$

Hence, the modified equation (1.1) is of the form

$$\begin{pmatrix} \dot{\widetilde{p}} \\ \dot{\widetilde{q}} \end{pmatrix} = \begin{pmatrix} \sum_{\tau\in TP_p} \frac{h^{|\tau|-1}}{\sigma(\tau)}\, b(\tau)\, F(\tau)(\widetilde{p},\widetilde{q}) \\ \sum_{\tau\in TP_q} \frac{h^{|\tau|-1}}{\sigma(\tau)}\, b(\tau)\, F(\tau)(\widetilde{p},\widetilde{q}) \end{pmatrix}, \tag{10.4}$$

where $b(\tau) = 1$ for $|\tau| = 1$, $b(\tau) = a(\tau) - \frac{1}{2}$ for $|\tau| = 2$. For $|\tau| > 2$, the coefficients $b(\tau)$ can be obtained recursively from Theorem 10.2 below. The proofs of the following two results are straightforward extensions of those for Lemma 9.1 and Theorem 9.2, and are therefore omitted.

Lemma 10.1 (Lie-Derivative of P-series). *Let $b(\tau)$ (with $b(\emptyset_p) = b(\emptyset_q) = 0$) and $c(\tau)$ be the coefficients of two P-series, and let $\big(p(t), q(t)\big)$ be a formal solution of the differential equation $h(\dot{p}(t), \dot{q}(t))^T = P\big(b, (p(t), q(t))\big)$, i.e., (10.4). The Lie derivative of the function $P\big(c, (p,q)\big)$ with respect to the vector field $P\big(b, (p,q)\big)$ is again a P-series*

$$h\,\frac{d}{dt}\, P\big(c, (p(t), q(t))\big) = P\big(\partial_b c, (p(t), q(t))\big).$$

Its coefficients are given by $\partial_b c(\emptyset_p) = \partial_b c(\emptyset_q) = 0$, and for $|\tau| \geq 1$ by

$$\partial_b c(\tau) = \sum_{\theta\in SP(\tau)} c(\theta)\, b(\tau\setminus\theta), \tag{10.5}$$

where, analogously to (9.5), $SP(\tau)$ denotes the set of splittings of $\tau \in TP$. □

In formula (10.5), $\emptyset_p \in SP(\tau)$ defines a splitting only if $\tau \in TP_p$, and $\emptyset_q \in SP(\tau)$ only if $\tau \in TP_q$. We therefore have $\partial_b c(\bullet) = c(\emptyset_p) b(\bullet)$, $\partial_b c(\circ) = c(\emptyset_q) b(\circ)$, and as examples for trees of order 3

$$\begin{aligned} \partial_b c(\substack{\circ\,\circ\\ \bullet}) &= c(\emptyset_p) b(\substack{\circ\,\circ\\ \bullet}) + 2\, c(\substack{\circ\\ \bullet}) b(\circ), \\ \partial_b c(\substack{\bullet\,\circ\\ \bullet}) &= c(\emptyset_p) b(\substack{\bullet\,\circ\\ \bullet}) + c(\substack{\bullet\\ \bullet}) b(\circ) + c(\substack{\circ\\ \bullet}) b(\bullet). \end{aligned}$$

Theorem 10.2. *If the method $(p_1, q_1) = \Phi_h(p_0, q_0)$ can be written as (10.2), the modified differential equation is given by (10.4), where the real coefficients $b(\tau)$ are recursively defined by $b(\emptyset_p) = b(\emptyset_q) = 0$, $b(\tau) = 1$ for $|\tau| = 1$, and*

$$b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|} \frac{1}{j!}\, \partial_b^{j-1} b(\tau) \qquad \text{for} \quad \tau \in TP. \tag{10.6}$$

Here, ∂_b^{j-1} denotes the iterate of the Lie derivative ∂_b defined in Lemma 10.1. □

Example 10.3. The symplectic Euler method

$$p_{n+1} = p_n + hf(p_{n+1}, q_n), \qquad q_{n+1} = q_n + hg(p_{n+1}, q_n) \tag{10.7}$$

is a partitioned Runge–Kutta method ($a_{11} = 1$, $\widehat{a}_{11} = 0$, $b_1 = \widehat{b}_1 = 1$) and can therefore be expressed as a P-series (10.2). From Theorem III.2.4 we get its coefficients:

$$a(\tau) = \begin{cases} 1 & \text{if all vertices (different from the root) are black,} \\ 0 & \text{otherwise.} \end{cases}$$

From Theorem 10.2 we can compute the coefficients $b(\tau)$ of the modified equation (10.4). They are given in Table 10.1 for the trees with a black root. Since $a(\tau)$ does not depend on the colour of the root of τ, the same holds for the coefficients $b(\tau)$. Hence, we do not include the values of $b(\tau)$ for trees with a white root.

Table 10.1. Coefficients $b(\tau)$ of the modified equation for symplectic Euler (10.7)

τ	$\bullet$									
$b(\tau)$	1	$1/2$	$-1/2$	$1/6$	$-1/3$	$1/6$	$1/3$	$-1/6$	$-1/6$	$1/3$

We know from Theorem 3.1 that the modified differential equation (10.4) of a symplectic method applied to a Hamiltonian system

$$\dot{p} = -H_q(p,q), \qquad \dot{q} = H_p(p,q) \tag{10.8}$$

is again Hamiltonian.

Theorem 10.4. *Suppose that for all separable Hamiltonians $H(p,q) = T(p) + U(q)$ the modified vector field (10.4), truncated after an arbitrary power of h, is (locally) Hamiltonian. Then, we have*

$$b(u \circ v) + b(v \circ u) = 0 \qquad u \in TP_p,\ v \in TP_q \tag{10.9}$$

for trees, where neighbouring vertices have different colours.

If it is (locally) Hamiltonian for all $H(p,q)$, then (10.9) holds for all $u \in TP_p$, $v \in TP_q$, and additionally we have

$$b(\tau) \text{ is independent of the colour of the root of } \tau \in TP. \tag{10.10}$$

If it is (locally) Hamiltonian for all $H(p,q) = \frac{1}{2}p^TCp + c^Tp + U(q)$ (with symmetric matrix C), then we have

$$b(\circ \circ u) + b(u \circ \circ) = 0, \quad b(u \circ\circ v) - b(v \circ\circ u) = 0 \qquad u,v \in TN_p \tag{10.11}$$

(see Sect. VI.7.1 for the definition of TN_p and $u \circ\circ v$).

The *proof* is the same as for Theorem 9.3 and therefore omitted. □

IX.10.2 Elementary Hamiltonians

We have already seen in Example 3.4 that the modified Hamiltonian of the symplectic Euler method is composed of expressions such as $H_p H_q$, $H_{pp}(H_q, H_q)$, $H_{pq}(H_q, H_p)$, etc. These will play the role of elementary Hamiltonians for partitioned methods. In the following definition, the elementary differentials $F(\tau)(p,q)$ correspond to the partitioned system $f(p,q) = -H_q(p,q)$, $g(p,q) = H_p(p,q)$.

Definition 10.5. For a given function $H : D \to \mathbb{R}$ (with open $D \subset \mathbb{R}^d \times \mathbb{R}^d$) and for $\tau \in TP$ we define the *elementary Hamiltonian* $H(\tau) : D \to \mathbb{R}$ by

$$\begin{aligned} H(\bullet)(p,q) &= H(\circ)(p,q) = H(p,q) \\ H(\tau)(p,q) &= \frac{\partial^{m+l} H(p,q)}{\partial^m p\, \partial^l q}\Big(F(u_1)(p,q), ..., F(v_1)(p,q), \ldots\Big) \end{aligned}$$

where $\tau = [u_1, \ldots, u_m, v_1, \ldots, v_l]_p$ or $\tau = [u_1, \ldots, u_m, v_1, \ldots, v_l]_q$ with trees $u_i \in TP_p$ and $v_i \in TP_q$.

Examples of elementary Hamiltonians are

$$H(\bullet) = H, \qquad H([\circ]_\bullet) = H_q H_p,$$
$$H([\bullet,\bullet]_\bullet) = H_{pp}(H_q, H_q), \quad H([\bullet,\circ]_\bullet) = -H_{pq}(H_q, H_p), \quad H([\circ,\circ]_\bullet) = H_{qq}(H_p, H_p).$$

We notice that, in contrast to Sect. IX.9.2, non-vanishing elementary Hamiltonians exist for trees with two vertices.

Lemma 10.6. *Elementary Hamiltonians satisfy*

$$H(u \circ v)(p,q) + H(v \circ u)(p,q) = 0 \qquad \textit{for } u \in TP_p \textit{ and } v \in TP_q, \tag{10.12}$$

and they do not depend on the colour of the root.

Proof. The independence of the colour of the root is by definition, and formula (10.12) is proved in the same way as the statement of Lemma 9.6. □

The conditions (10.9) and (10.10) define relations between the coefficients $b(\tau)$ of a Hamiltonian vector field (10.4). The previous lemma shows analogous relations between elementary Hamiltonians. This motivates the consideration of the following equivalence relation on TP (Hairer 1994).

Definition 10.7. We denote by $\sim$ the smallest *equivalence relation* on TP which satisfies the two properties

- $u \sim v$ if u and v are identical with the exception of the colour of the root;
- $u \circ v \sim v \circ u$ for $u \in TP_p$ and $v \in TP_q$.

Fig. 10.1. Groups of equivalent trees of orders up to three

Equivalent trees of orders up to three are grouped together in Fig. 10.1. We can change the colour of the root, and we can move the root to a neighbouring vertex if it has the opposite colour.

In the case of separable Hamiltonians, one has to consider only trees for which neighbouring vertices have different colours. This implies that the first condition of Definition 10.7 is empty. The second condition means that the root can be moved arbitrarily in the tree without changing the equivalence class. For this special situation, equivalence classes have been considered already by Abia & Sanz-Serna (1993) and are named "bicolour (unrooted) trees".

Similar to (9.14) we select representatives from the equivalence class as follows: we fix a total ordering on the set TP that (i) respects the number of vertices, and (ii) is such that no tree is between trees that differ only in the colour of the root. The ordering of Fig. 10.1 is such a possible choice. We then define

$$TP^* = \Big\{ \bullet, \circ \Big\} \cup \Big\{ \tau \in TP \;\Big|\; \begin{array}{l} \tau \text{ cannot be written as } \tau = u \circ v \text{ with } u < v, \\ \text{also not if the colour of the root is changed.} \end{array} \Big\}. \tag{10.13}$$

We further let $TP_p^* = TP^* \cap TP_p$ and $TP_q^* = TP^* \cap TP_q$.

Lemma 10.8. *For a tree $\tau \in TP^*$ we have*

$$\begin{aligned} -\frac{\partial H(\tau)}{\partial q}(p,q) &= \sigma(\tau) \sum_{\theta\sim\tau,\theta\in TP_p} \frac{(-1)^{\kappa(\tau,\theta)}}{\sigma(\theta)} F(\theta)(p,q), \\ \frac{\partial H(\tau)}{\partial p}(p,q) &= \sigma(\tau) \sum_{\theta\sim\tau,\theta\in TP_q} \frac{(-1)^{\kappa(\tau,\theta)}}{\sigma(\theta)} F(\theta)(p,q), \end{aligned} \tag{10.14}$$

where $\kappa(\tau,\theta)$ is the number of root changes that are necessary to obtain θ from τ.

The *proof* is the same as for Lemma 9.7 and therefore omitted. □

We are now able to give the main result of this section.

Theorem 10.9. *Consider a numerical method that can be written as a P-series (10.2), and that is symplectic for every Hamiltonian (10.8). Its modified differential equation is then Hamiltonian with*

$$\widetilde{H}(p,q) = H_1(p,q) + h\, H_2(p,q) + h^2\, H_3(p,q) + \ldots,$$

where

$$H_j(p,q) = \sum_{\tau\in TP_p^*,\, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)}\, H(\tau)(p,q), \tag{10.15}$$

and the coefficients $b(\tau)$ are those of Theorem 10.2. Notice that $H_j(p,q)$ from (10.15) is independent of whether we sum over trees in TP_p^* or TP_q^*.

Proof. This is the same as for Theorem 9.8. □

If the method (10.2) is known to be symplectic for separable Hamiltonians only, and if it is applied to $H(p,q) = T(p) + U(q)$, the statement of Theorem 10.9 is still valid. In this situation $H(\tau)(p,q)$ vanishes if a vertex of τ has sons with different colour (it then contains a factor $H_{pq\ldots} = 0$).

Example 10.10. Consider the 2-stage Lobatto IIIA - IIIB pair (cf. Table II.2.1), which is the natural extension of the Störmer–Verlet scheme to non-separable problems. We compute the coefficients $a(\tau)$ from Theorem III.2.4, and $b(\tau)$ from Theorem 10.2. The result is given in Table 10.2. Notice that $a(\tau)$ and $b(\tau)$ are both independent of the colour of the root. Theorem 10.9 then yields

$$\widetilde{H} = H + \frac{h^2}{24}\Big(2H_{pp}H_q^2 - H_{qq}H_p^2 + 2H_{pq}H_qH_p\Big) + \ldots \tag{10.16}$$

for the modified Hamiltonian. Since the method is symmetric, $\widetilde{H}$ is in even powers of h. The next non-vanishing term requires the consideration of trees up to order 5.

Table 10.2. Coefficients $a(\tau)$ and $b(\tau)$ for the Störmer–Verlet scheme (Table II.2.1)

τ										
$a(\tau)$	1	$1/2$	$1/2$	$1/2$	$1/4$	$1/4$	$1/4$	$1/4$	0	$1/4$
$b(\tau)$	1	0	0	$1/6$	$-1/12$	$-1/12$	$1/12$	$1/12$	$-1/6$	$1/12$

Remark 9.9, the characterization of symplectic vector fields (10.4), and the results of Sect. IX.9.4 can be extended to the case of (partitioned) P-series. We renounce of giving all the details here.

IX.11 Exercises

1. Change the Maple program of Example 1.1 in such a way that the modified equations for the implicit Euler method, the implicit midpoint rule, or the trapezoidal rule are obtained. Observe that for symmetric methods one gets expansions in even powers of h.
2. Write a short Maple program which, for simple methods such as the symplectic Euler method, computes some terms of the modified equation for a two-dimensional system $\dot{p} = f(p,q)$, $\dot{q} = g(p,q)$. Check the modified equations of Example 1.3.
3. Prove that the modified equation of the Störmer–Verlet scheme (I.1.15) applied to $\ddot{y} = g(y)$ is a second order differential equation of the form $\ddot{\widetilde{y}} = g_h(\widetilde{y}, \dot{\widetilde{y}})$ with initial values given by $\widetilde{y}(0) = y_0$ and $\dot{\widetilde{y}}(0)$ such that $\widetilde{y}(h) = y_1$ holds.

Hint. Taylor expansion shows that for a smooth function $\widetilde{y}(t)$ satisfying $\widetilde{y}(t) = y_n$ we have

$$\Big(1 + \frac{h^2}{12}\, D^2 + \frac{h^4}{360}\, D^4 + \ldots\Big)\ddot{\widetilde{y}}(t) = g\big(\widetilde{y}(t)\big),$$

where D represents differentiation with respect to time.
Warning. In general, we do not have that $\dot{\widetilde{y}}(t_n) = \dot{y}_n$.

4. Prove that for ρ-reversible differential equations the elementary differentials satisfy
$$F(\tau)(\rho y) = (-1)^{|\tau|}\rho\, F(\tau)(y).$$
Use this to give an alternative proof of Theorem 2.3 for the case that the method is symmetric and can be expressed as a B-series.
5. Find a first integral of the truncated modified equation for the symplectic Euler method and the Lotka–Volterra problem (Example 1.3).
Hint. With the transformation $p = \exp P$, $q = \exp Q$ you will get a Hamiltonian system.
Result. $\widetilde{I}(p,q) = I(p,q) - h\big((p+q)^2 - 8p - 10q + 2\ln p + 8\ln q\big)/4$.
6. (Field & Nijhoff 2003). Apply the symplectic Euler method to the system with Hamiltonian $H(p,q) = \ln(\alpha + p) + \ln(\beta + q)$. Compute the modified Hamiltonian and prove that the series converges for sufficiently small step sizes.
Hint. The method conserves exactly $I(p,q) = (\alpha+p)(\beta+q)$. Find linear two-term recursions for $\{p_n\}$ and $\{q_n\}$, and use the ideas of Example 1.4. *Result.*
$$\widetilde{H}(p,q) = H(p,q) - \sum_{k\ge 1} \frac{h^k\, I(p,q)^{-k}}{k(k+1)}.$$
7. Compute $\partial_b c(\tau)$ for the tree $\tau = [[\tau],\tau]$ of order 4.
8. For the implicit midpoint rule compute the coefficients $a(\tau)$ of the expansion (9.1), and also a few coefficients $b(\tau)$ of the modified equation.
Result. $a(\tau) = 2^{1-|\tau|}$, $b(\bullet) = 1$, $b([\bullet]) = 0$, $b(\tau) = a(\tau) - 1/\gamma(\tau)$ for $|\tau| = 3$.
9. Check the formulas of Table 9.1.
10. Consider a differential equation $\dot{y} = f(y)$ with a divergence-free vector field, and apply a volume-preserving integrator. Show that every truncation of the modified equation has again a divergence-free vector field.
Hint. Adapt the proof by induction of Theorems 2.3 and 3.1.
11. Consider explicit 2-stage Runge–Kutta methods of order 2, applied to the pendulum problem $\dot{q} = p$, $\dot{p} = -\sin q$. With the help of Exercise 2 compute $f_3(p,q)$ of the modified differential equation. Is there a choice of the free parameter c_2, such that $f_3(p,q)$ is a Hamiltonian vector field?
12. Find at least two linear transformations ρ for which the Kepler problem (I.2.2), written as a first order system, is ρ-reversible.
13. Consider the Kepler problem (I.2.2), written as a Hamiltonian system (I.1.10). Find constants M and R such that (7.2) holds for all $(p,q) \in \mathbb{R}^4$ satisfying
$$\|p\| \le 2 \qquad \text{and} \qquad 0.8 \le \|q\| \le 1.2.$$

14. (McLachlan & Zanna 2005). Consider the RATTLE method (Algorithm VII.5.1) applied to the Euler equations (VII.5.10) of the free rigid body, written as $\dot{y} = f(y)$. Prove that the modified differential equation is of the form

$$\dot{y} = (1 + h^2 s_2(y) + h^4 s_4(y) + \ldots)\, f(y), \tag{11.1}$$

where the scalar functions $s_k(y)$ depend on y only via the Casimir function $C(y) = y_1^2 + y_2^2 + y_3^2$ and the Hamiltonian $H(y) = \frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$. Consequently, all $s_k(y)$ are constant along solutions of the Euler equations.
Hint. Since $C(y)$ and $H(y)$ are exactly conserved by the numerical method (see Sect. VII.5.3), the modified equation is a time transformation of the original system. The special form of the functions $s_k(y)$ follows from the fact that RATTLE is a Poisson integrator (Theorem VII.5.11) and from a transformation to canonical form as in Theorem 3.5.
15. (Murua 1999). Let $\Phi_h(y) = B(a, y)$ be given by a B-series and denote with $b(\tau)$ the coefficients of the corresponding modified differential equation, cf. formula (9.4). Prove that the coefficients of the nth iterate $\Phi_h^n(y) = B(a^n, y)$ satisfy
$$a^n(\tau) = n\, b(\tau) + n^2 c(\tau, n) \qquad \text{for } \tau \in T,$$
where $c(\tau, n)$ is a polynomial of degree $|\tau| - 2$ in n.
Hint. This follows from the Taylor series $\widetilde{y}(nh) = \widetilde{y}(0) + nh\widetilde{y}'(0) + \ldots$ for the solution of the modified differential equation.
16. With the help of Exercise 15, give an alternative proof of Theorem 9.3.
Hint. If $B(a, y)$ is symplectic, also $B(a^n, y)$ is symplectic and its coefficients thus satisfy (VI.7.4).
17. (Murua 1997). Find a one-to-one correspondence between the equivalence classes of TP (corresponding to $\sim$ of Definition 10.7) and *oriented free trees* (i.e., trees without a distinguished vertex (root), but with oriented edges), see Fig. 11.1.

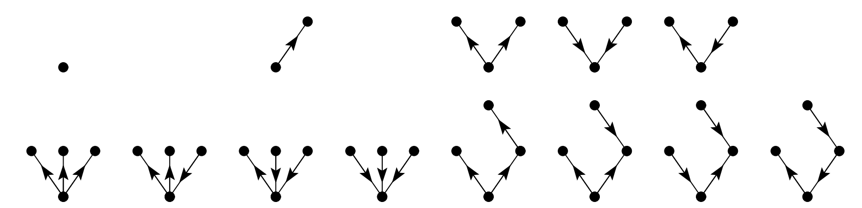


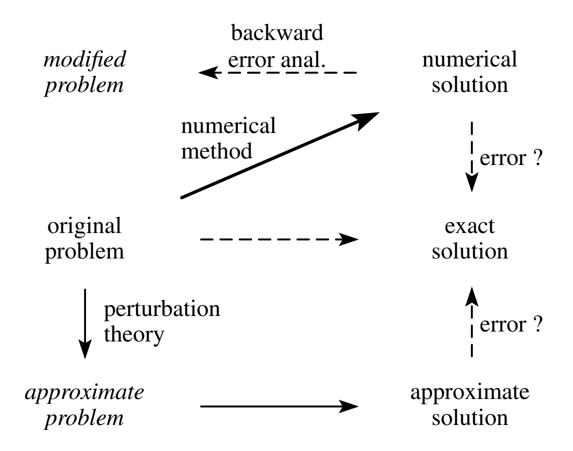
Fig. 11.1. Oriented free trees up to order four

Chapter X. Hamiltonian Perturbation Theory and Symplectic Integrators

> Perturbation theory is in fact an outgrowth of the necessity to determine the orbits with ever greater accuracy. This problem can be solved today, but in what is for the theoretician a rather disappointing way. With modern calculating machines, one is now able to compute directly results even more accurately than those provided by perturbation theory.
> (J. Moser 1978)

> . . . allows computer prediction of planetary positions far more accurate (by brute computation) than anything provided by classical perturbation theory. In a very real sense, one of the most exhalted of human endeavors, going back to the priests of Babylon and before, has been taken over by the machine. (S. Sternberg 1969)

In this chapter we study the long-time behaviour of symplectic integrators, combining backward error analysis and the perturbation theory of integrable Hamiltonian systems.



During the 18th and 19th centuries, scientists struggled for the integration of complicated problems of dynamics, with the main aim of solving them analytically by "quadrature". But only few problems could be treated successfully in this way. In cases where the original problem could not be solved, much effort was put into re-

placing it by an integrable *approximate problem*, by using and developing perturbation theory. Thereby, a rich arsenal of very ingenuous theories has been discovered since the 19th century.

In the 1960s and 1970s, the enormous progress of “calculating machines” and numerical software allowed many of the original problems to be solved with extreme accuracy, so that for the first time numerical integration methods superseded analytical perturbation methods in the computations of celestial mechanics (see the above citations). Since then, the further increase in computing speed has allowed problems to be treated on larger and larger time scales, where huge amounts of errors are accumulated and need to be understood and controlled. In the spirit of backward error analysis, these numerical errors are interpreted as those of a *modified problem*, for the study of which perturbation theory is once again the appropriate tool.

X.1 Completely Integrable Hamiltonian Systems

Integrable Hamiltonian systems were originally of interest because their equations of motion can be solved analytically. Their interest in the present context lies in the fact that their flow is simply uniform motion on a Cartesian product of circles and straight lines in suitable coordinates, and that many physical systems can be viewed as perturbations of integrable systems.

X.1.1 Local Integration by Quadrature

> M. Liouville a fait voir qu’il fallait que toutes les combinaisons (α, β) des intégrales trouvées fussent nulles. (E. Bour 1855)

One of the great dreams of 18th and 19th century analytical mechanics was to solve the equations of motion of mechanical systems by “quadrature”, that is, using only evaluations and inversions of functions and calculating integrals of known functions. In this spirit, Newton’s (1687) equations of motion of Kepler’s two-body problem were solved by Joh. Bernoulli (1710) and Newton (1713), see Sect. I.2.2. Euler’s (1760) solution of the problem of the attraction of a particle by two fixed centres, and Lagrange’s (1766) study of motion of a particle in a field with one attracting centre and under an additional constant force were among the important achievements of the 18th century. The three-body problem, however, resisted all efforts aiming at an integration by quadrature, and though it continued to do so, this problem spurred the development of extremely useful mathematical theories of a much wider scope throughout the 19th century, from Poisson to Poincaré via Hamilton, Jacobi, Liouville, to name but a few of the most eminent mathematicians contributing to analytical mechanics.

Consider the Hamiltonian system

$$\dot{p} = -\frac{\partial H}{\partial q}(p, q) , \quad \dot{q} = \frac{\partial H}{\partial p}(p, q) , \tag{1.1}$$

with d degrees of freedom: $(p, q) \in \mathbb{R}^d \times \mathbb{R}^d$. We try to find a symplectic transformation $(p, q) \mapsto (x, y)$, such that the system has a more amenable form in the new coordinates. In particular, this is the case if the Hamiltonian expressed in the new variables,

$$H(p, q) = K(x) , \tag{1.2}$$

does not depend on y. Since $\frac{\partial K}{\partial y} \equiv 0$, the transformed system then becomes (recall the conservation of the Hamiltonian form of the differential equations under symplectic transformations, Theorem VI.2.8)

$$\dot{x} = 0 , \quad \dot{y} = \omega(x) , \tag{1.3}$$

with $\omega(x) = \frac{\partial K}{\partial x}(x)$. This is readily integrated:

$$x(t) = x_0 , \quad y(t) = y_0 + \omega(x_0)t .$$

As we recall from Sect. VI.5, a symplectic transformation $(p, q) \mapsto (x, y)$ can be constructed via a *generating function* $S(x, q)$ by the equations

$$y = \frac{\partial S}{\partial x}(x, q) , \quad p = \frac{\partial S}{\partial q}(x, q) . \tag{1.4}$$

If (p_0, q_0) and (x_0, y_0) are related by (1.4), and if $\partial^2 S/\partial x \partial q$ is invertible at (x_0, q_0), then the equations (1.4) define a symplectic transformation between neighbourhoods of (p_0, q_0) and (x_0, y_0).

The equation (1.2) together with the second equation of (1.4) give a partial differential equation for S, the *Hamilton–Jacobi equation*

$$H\Big(\frac{\partial S}{\partial q}(x, q), q\Big) = K(x) .$$

If $S(x, q)$ is a solution of such an equation (for some function K), then (1.3) shows that $x_i = F_i(p, q)$ ($i = 1, \ldots, d$) as given implicitly by the second equation of (1.4), are first integrals of the Hamiltonian system (1.1). Moreover, these functions F_i are *in involution*, which means that their Poisson brackets vanish pairwise:

$$\{F_i, F_j\} = 0, \quad i, j = 1, \ldots, d .$$

This is an immediate consequence of the definition $\{F, G\} = \nabla F^T J^{-1} \nabla G$ of the Poisson bracket and of the symplecticity of the transformation (the left upper block of J^{-1} is 0).

Conversely, it was realized by Bour (1855) and Liouville (1855) that a Hamiltonian system having d first integrals in involution can *locally* be transformed to the form (1.3) by "quadrature". This observation is based on the following completion result and its proof.

Lemma 1.1 (Liouville Lemma). *Let $F_1, \dots, F_d$ be smooth real-valued functions, defined in a neighbourhood of $(p_0, q_0) \in \mathbb{R}^d \times \mathbb{R}^d$. Suppose that these functions are in involution (i.e., all Poisson brackets $\{F_i, F_j\} = 0$), and that their gradients are linearly independent at (p_0, q_0). Then, there exist smooth functions $G_1, \dots, G_d$, defined on some neighbourhood of (p_0, q_0), such that*

$$(F_1, \dots, F_d, G_1, \dots, G_d) : (p, q) \mapsto (x, y) \quad \textit{is a symplectic transformation.}$$

Proof. Let $F = (F_1, \dots, F_d)^T$. The linear independence of the gradients ∇F_i implies that there are d columns of the $d \times 2d$ Jacobian $\partial F / \partial(p, q)$ that form an invertible $d \times d$ submatrix. After some suitable symplectic transformations (see Exercise 1) we may assume without loss of generality that $F_p = \partial F / \partial p$ is invertible. By the implicit function theorem, we can then locally solve $x = F(p, q)$ for p:

$$p = P(x, q) \quad \text{with partial derivatives} \quad P_x = F_p^{-1}, \; P_q = -F_p^{-1} F_q \; .$$

The condition that the F_i are in involution, reads in matrix notation

$$F_p F_q^T - F_q F_p^T = 0 \; .$$

Multiplying this equation with F_p^{-1} from the left and with F_p^{-T} from the right, we obtain

$$-P_q^T + P_q = 0 \; ,$$

so that $P_q = \partial P / \partial q$ is symmetric. By the Integrability Lemma VI.2.7, $P(x, q)$ is thus locally the gradient with respect to q of some function $S(x, q)$ (which is constructed by quadrature). Moreover, $\frac{\partial^2 S}{\partial x \partial q} = P_x = F_p^{-1}$ is invertible. The equations (1.4) define a symplectic transformation $(p, q) \mapsto (x, y)$, and by construction $x = F(p, q)$. □

If, in a Hamiltonian system with d degrees of freedom, we can find d independent first integrals in involution $H = F_1, F_2, \dots, F_d$, then Lemma 1.1 yields a symplectic change of coordinates, constructed by quadrature, which transforms (1.1) *locally* to (1.2) with $K(x_1, \dots, x_d) = x_1$.

Example 1.2. Consider the Hamiltonian of motion in a central field,

$$H = \tfrac{1}{2}(p_1^2 + p_2^2) + V(r) \qquad \text{for} \quad r = \sqrt{q_1^2 + q_2^2} \; ,$$

with a potential $V(r)$ that is defined and smooth for $r > 0$. The Kepler problem corresponds to the special case $V(r) = -1/r$, and the perturbed Kepler problem to $V(r) = -1/r - \mu/(3r^3)$. Changing to polar coordinates (see Example VI.5.2)

$$\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} , \qquad \begin{pmatrix} p_r \\ p_\varphi \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -r\sin\varphi & r\cos\varphi \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} , \tag{1.5}$$

this becomes

$$H(p_r, p_\varphi, r, \varphi) = \frac{1}{2}\Big(p_r^2 + \frac{p_\varphi^2}{r^2}\Big) + V(r) \; .$$

The system has the angular momentum $L = p_\varphi$ as a first integral, since H does not depend on φ. Clearly, $\{H, L\} = 0$ everywhere. The gradients of H and L are linearly independent unless both $p_r = 0$ and $p_\varphi^2 = r^3 V'(r)$. By inserting $p_\varphi^2 = 2r^2(H - V(r))$ and eliminating r this becomes a condition of the form $\alpha(H, L) = 0$, which for the Kepler problem reads explicitly $L^2(1 + 2HL^2) = 0$. The conditions of Lemma 1.1 are thus satisfied on the domain

$$M = \{(p_r, p_\varphi, r, \varphi) \; ; \; r > 0, \; \alpha(H, L) \neq 0\} \; .$$

The equations $x_1 = H = \frac{1}{2}(p_r^2 + p_\varphi^2/r^2) + V(r)$, $x_2 = L = p_\varphi$ can be solved for

$$p_r = \pm\sqrt{2(H - V(r)) - L^2/r^2} \; , \quad p_\varphi = L \; ,$$

and $p_r = \partial S/\partial r$, $p_\varphi = \partial S/\partial\varphi$ with

$$S(H, L, r, \varphi) = L\varphi \pm \int_{r_0}^{r} \sqrt{2(H - V(\rho)) - L^2/\rho^2} \, d\rho \; .$$

The conjugate variables are

$$\begin{aligned} y_1 &= \frac{\partial S}{\partial H} = \pm \int_{r_0}^{r} \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} \, d\rho \; , \\ y_2 &= \frac{\partial S}{\partial L} = \varphi \mp \int_{r_0}^{r} \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} \, d\rho \; . \end{aligned} \tag{1.6}$$

This defines (locally) the transformation $(p_r, p_\varphi, r, \varphi) \mapsto (x_1, x_2, y_1, y_2)$. In these variables, the equations of motion read $\dot{x}_1 = 0$, $\dot{x}_2 = 0$, $\dot{y}_1 = 1$, $\dot{y}_2 = 0$. Over any time interval where $p_r(t)$ does not change sign, solutions therefore satisfy

$$\begin{aligned} t_1 - t_0 &= \pm \int_{r(t_0)}^{r(t_1)} \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} \, d\rho \; , \\ \varphi(t_1) - \varphi(t_0) &= \pm \int_{r(t_0)}^{r(t_1)} \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} \, d\rho \; . \end{aligned} \tag{1.7}$$

X.1.2 Completely Integrable Systems

Lemma 1.1 appears as a powerful tool for an explicit solution by quadrature. However, because of its purely local nature this lemma does not tell us anything about the dynamics of the system. This was not a concern at Liouville's time, but the first rigorous non-integrability results by Poincaré (1892) put a definite end to the hope of being eventually able to construct explicit analytic solutions of most equations of motion by quadrature, and shifted the interest to understanding the *global*, qualitative behaviour of dynamical systems.

Lemma 1.1 can be globalized by a procedure similar to analytic continuation if the conditions of the following definition are satisfied.

Definition 1.3. A Hamiltonian system with Hamiltonian $H : M \to \mathbb{R}$ (M an open subset of $\mathbb{R}^d \times \mathbb{R}^d$) is called *completely integrable* if there exist smooth functions $F_1 = H, F_2, \ldots, F_d : M \to \mathbb{R}$ with the following properties:
1) $F_1, \ldots, F_d$ are in involution (i.e., all $\{F_i, F_j\} = 0$) on M.
2) The gradients of $F_1, \ldots, F_d$ are linearly independent at every point of M.
3) The solution trajectories of the Hamiltonian systems with Hamiltonian F_i ($i = 1, \ldots, d$) exist for all times and remain in M.

Obviously, all the Hamiltonian systems with Hamiltonian F_i ($i = 1, \ldots, d$) are then completely integrable, and so there will be no mathematical reason to further distinguish $H = F_1$. We note that condition (1) of Definition 1.3 implies that all F_j are first integrals of the Hamiltonian system with Hamiltonian F_i, and that the flows $\varphi_t^{[i]}$ of these Hamiltonian systems commute: $\varphi_t^{[i]} \circ \varphi_s^{[j]} = \varphi_s^{[j]} \circ \varphi_t^{[i]}$ for all i, j and all $t, s \in \mathbb{R}$; see Lemma VII.3.2.

For $x = (x_i) \in \mathbb{R}^d$ we define the level set

$$M_x = \{(p, q) \in M \;;\; F_i(p, q) = x_i \;\text{ for } i = 1, \ldots, d\}. \tag{1.8}$$

Theorem 1.4. *Suppose that $F_1, \ldots, F_d : M \to \mathbb{R}$ satisfy the conditions of Definition 1.3. Assume that M_x is connected (and non-empty) for all x in a neighbourhood of $x_0 \in \mathbb{R}^d$. Then, on some neighbourhood B of x_0, there exists a symplectic and surjective mapping*

$$e : B \times \mathbb{R}^d \to \bigcup_{x \in B} M_x : (x, y) \mapsto (p, q) \in M_x$$

that linearizes, for all $i = 1, \ldots, d$, the flow $\varphi_t^{[i]}$ of the system with Hamiltonian F_i:

$$\textit{if} \quad (p, q) = e(x, y), \quad \textit{then} \quad \varphi_t^{[i]}(p, q) = e(x, y + t e_i), \tag{1.9}$$

where $e_i = (0, \ldots, 1, \ldots, 0)^T$ is the ith unit vector of $\mathbb{R}^d$.

Since e is symplectic, e is a local diffeomorphism. Its local inverse is a transformation as constructed in Lemma 1.1. However, (p, q) can have countably many discretely lying pre-images (x, y), so that e^{-1} becomes a multi-valued function. The situation is analogous to that of the complex exponential and logarithm. The following example illustrates that this analogy is not incidental.

Example 1.5. Consider the harmonic oscillator, i.e., $d = 1$ and $H(p, q) = \frac{1}{2}(p^2 + q^2)$. For $x = \frac{1}{2}r^2$, we have $e(x, y) = (r \cos y, r \sin y)$.

Proof of Theorem 1.4. We fix $(p_0, q_0) \in M_{x_0}$, and in a neighbourhood U of (p_0, q_0) we consider a symplectic transformation

$$\ell = (F_1, \ldots, F_d, G_1, \ldots, G_d) : (p, q) \mapsto (x, y)$$

as constructed in Lemma 1.1. We have $\ell(p_0, q_0) = (x_0, y_0)$ where we may assume $y_0 = 0$. To every $v = (v_i) \in \mathbb{R}^d$ we associate the Hamiltonian

$$F_v = v_1 F_1 + \ldots + v_d F_d$$

and note that, because of the commutativity of the flows $\varphi_t^{[i]}$, the flow of the system with Hamiltonian F_v equals

$$\varphi_{tv} = \varphi_{tv_1}^{[1]} \circ \ldots \circ \varphi_{tv_d}^{[d]} .$$

In the neighbourhood U of (p_0, q_0), the system with Hamiltonian F_v is transformed under the symplectic mapping ℓ to

$$\dot{x} = 0, \quad \dot{y} = v .$$

Hence, the following diagram commutes for $(p,q) \in U$ and for sufficiently small tv:

$$\begin{array}{ccc} (p,q) & \longrightarrow & \varphi_{tv}(p,q) \\ \Big\downarrow \ell & & \Big\uparrow \ell^{-1} \\ (x,y) & \longrightarrow & (x, y+tv) \end{array} \tag{1.10}$$

We now construct e by extending this diagram to arbitrary tv:

$$\begin{array}{ccc} (p,q) & \longrightarrow & \varphi_{y}(p,q) \\ \Big\uparrow \ell^{-1} & & \\ (x,0) & \longleftarrow & (x,y) \end{array} \tag{1.11}$$

That is, we define on $B \times \mathbb{R}^d$ (with B a neighbourhood of x_0 on which $\ell^{-1}(x,0)$ is defined)

$$e(x,y) = \varphi_y\big(\ell^{-1}(x,0)\big) .$$

For (x,y) near some fixed $(\widehat{x}, \widehat{y})$, we have by (1.10) with $y - \widehat{y}$ and $\widehat{y}$ instead of y and tv that

$$e(x,y) = \varphi_{\widehat{y}}\big(\ell^{-1}(x, y - \widehat{y})\big) ,$$

which shows that e is symplectic, being locally the composition of symplectic transformations. The property (1.9) is obvious from the definition of e and from the commutativity of the flows $\varphi_t^{[i]}$. Since $\ell^{-1}(x,0) \in M_x$ and M_x is invariant under the flows $\varphi_t^{[i]}$, we have $e(x,y) \in M_x$ for all (x,y).

It remains to show that $e : \{x\} \times \mathbb{R}^d \to M_x$ is surjective for every x near x_0. Let $(\widehat{p}, \widehat{q})$ be an arbitrary point on M_x. By assumption, there exists a path on M_x connecting $\ell^{-1}(x,0)$ and $(\widehat{p}, \widehat{q})$. Moreover, by (1.10) and by the compactness of the path, there is a $\delta > 0$ such that, for every (p,q) on this path, the mapping $y \mapsto \varphi_y(p,q)$ is a diffeomorphism between the ball $\|y\| < \delta$ and a neighbourhood of (p,q) on M_x. Therefore, $(\widehat{p}, \widehat{q})$ can be reached from $\ell^{-1}(x,0)$ by a finite composition of maps:

$$(\widehat{p}, \widehat{q}) = \varphi_{y^{(m)}} \circ \ldots \circ \varphi_{y^{(1)}}(\ell^{-1}(x,0)) = \varphi_{\widehat{y}}(\ell^{-1}(x,0)) = e(x, \widehat{y}) ,$$

where $\widehat{y} = y^{(1)} + \ldots + y^{(m)}$ once again by the commutativity of the flows $\varphi_t^{[i]}$. $\square$

Illustration of the Liouville Transform. We illustrate the above construction at a simple example, the pendulum (I.1.12) with Hamiltonian $H = p^2/2 - \cos q$. The first coordinate is $x = H(p,q)$, a first integral. The second coordinate y is, following (1.11), the time t which is necessary to reach the point (p,q) from an initial line, which we assume at $q = 0$. Then we have (Fig. 1.1 left) $dp\,dq = dH\,dt$ (because

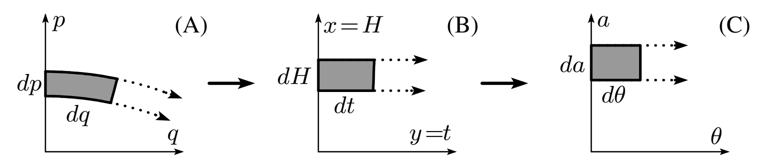


Fig. 1.1. Liouville and action-angle coordinate transforms

of $dq = H_p\,dt$ and $dH = H_p\,dp$). We see again that we have area preservation, because the symplecticity of the flow preserves this property for all times. This symplectic change of coordinates $(p,q) \mapsto (x,y)$ is illustrated in Fig. 1.2, which transforms the problem (A) to a much simpler form (B) with uniform horizontal movement.

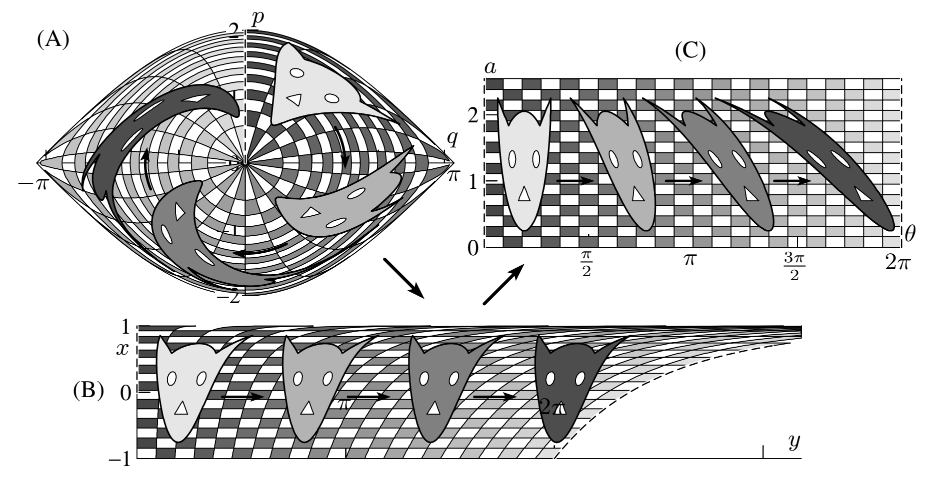


Fig. 1.2. Liouville and action-angle coordinates illustrated at the pendulum problem

We are not yet completely satisfied, however, because the orbits have periods $g = g(H)$ which are not all the same. We therefore append a *second* transform by putting $\theta = \frac{2\pi}{g} \cdot t$ (see picture (C) in Fig. 1.1 and Fig. 1.2), which forces all periods into a Procrustean bed of length 2π. Area preservation $da\,d\theta = dH\,dt$ now requires that $2\pi\,da = g(H)\,dH$, which is a differential equation between a and H. The new coordinates (a,θ) are the *action-angle variables* and we see that they transform the phase space into $D \times \mathbb{T}^1$ where $D \subset \mathbb{R}^1$. We again have horizontal movement, but this time the speed depends on a. The general existence for completely integrable systems will be proved in Theorem 1.6 below.

X.1.3 Action-Angle Variables

> We show here that, under the hypotheses of Liouville's theorem, we can find symplectic coordinates $(\mathbf{I}, \boldsymbol{\varphi})$ such that the first integrals $\mathbf{F}$ depend only on $\mathbf{I}$, and $\boldsymbol{\varphi}$ are angular coordinates on the torus $M_{\mathbf{f}}$.
>
> (V.I. Arnold 1989, p. 279)

We are now in the position to prove the main result of this section, which establishes a symplectic change of coordinates to the so-called *action-angle variables*, such that d first integrals of a completely integrable system depend only on the actions, and the angles are defined globally mod 2π (provided the level sets of the first integrals are compact). This is known as the Arnold–Liouville theorem; cf. Arnold (1963, 1989), Arnold, Kozlov & Neishtadt (1997; Ch. 4, Sect. 2.1), Jost (1968). Here and in the following,

$$\mathbb{T}^d = \mathbb{R}^d/2\pi\mathbb{Z}^d = \{(\theta_1 \bmod 2\pi, \dots, \theta_d \bmod 2\pi)\ ;\ \theta_i \in \mathbb{R}\}$$

denotes the standard d-dimensional torus.

Theorem 1.6 (Arnold–Liouville Theorem). *Let $F_1, \dots, F_d : M \to \mathbb{R}$ be first integrals of a completely integrable system as in Definition 1.3. Suppose that the level sets M_x (see (1.8)) are compact and connected for all x in a neighbourhood of $x_0 \in \mathbb{R}^d$. Then, there are neighbourhoods B of x_0 and D of 0 in $\mathbb{R}^d$ such that the following holds:*

(i) For every $x \in B$, the level set M_x is a d-dimensional torus that is invariant under the flow of the system with Hamiltonian F_i $(i = 1, \dots, d)$.

(ii) There exists a bijective symplectic transformation

$$\psi : D \times \mathbb{T}^d \to \bigcup_{x \in B} M_x \subset \mathbb{R}^d \times \mathbb{R}^d : (a, \theta) \mapsto (p, q)$$

such that $(F_i \circ \psi)(a, \theta)$ depends only on a, i.e.,

$$F_i(p, q) = f_i(a) \quad \textit{for} \quad (p, q) = \psi(a, \theta) \qquad (i = 1, \dots, d)$$

with functions $f_i : D \to \mathbb{R}$.

The variables $(a, \theta) = (a_1, \dots, a_d, \theta_1 \bmod 2\pi, \dots, \theta_d \bmod 2\pi)$ are called *action-angle variables*.

Remark 1.7. If the level sets M_x are not compact, then the proof of Theorem 1.6 shows that M_x is diffeomorphic to a Cartesian product of circles and straight lines $\mathbb{T}^k \times \mathbb{R}^{d-k}$ for some $k < d$, and there is a bijective symplectic transformation $(a, \theta) \mapsto (p, q)$ between $D \times (\mathbb{T}^k \times \mathbb{R}^{d-k})$ and a neighbourhood $\bigcup \{M_x : x \in B\}$ of M_{x_0} such that the first integrals again depend only on a.

Remark 1.8. If the Hamiltonian is real-analytic, then the proof shows that also the transformation to action-angle variables is real-analytic.

Proof of Theorem 1.6. (a) We return to Theorem 1.4. For $x \in B$, we consider the set

$$\Gamma_x = \{y \in \mathbb{R}^d \; ; \; e(x,y) = e(x,0)\} \; .$$

Since e is locally a diffeomorphism, for every fixed $y_0 \in \Gamma_{x_0}$ there exists a unique smooth function η defined on a neighbourhood of x_0, such that $\eta(x_0) = y_0$ and $\eta(x) \in \Gamma_x$ for x near x_0. In particular, Γ_x is a discrete subset of $\mathbb{R}^d$. By (1.9), for $y \in \Gamma_x$ we have $e(x, y+v) = e(x,v)$ for all $v \in \mathbb{R}^d$. Therefore, Γ_x is a subgroup of $\mathbb{R}^d$, i.e., with $y, v \in \Gamma_x$ also $y+v \in \Gamma_x$ and $-y \in \Gamma_x$. It then follows (see Exercise 4) that Γ_x is a grid, generated by $k \le d$ linearly independent vectors $g_1(x), \ldots, g_k(x) \in \mathbb{R}^d$:

$$\Gamma_x = \{m_1 g_1(x) + \ldots + m_k g_k(x) \; ; \; m_i \in \mathbb{Z}\} \; .$$

We extend $g_1(x), \ldots, g_k(x)$ to a basis $g_1(x), \ldots, g_d(x)$ of $\mathbb{R}^d$. Then, e induces a diffeomorphism

$$\begin{array}{rcl} \mathbb{T}^k \times \mathbb{R}^{d-k} & \to & M_x \\ (\theta_1, \ldots, \theta_k, \tau_{k+1}, \ldots, \tau_d) & \mapsto & e\Big(x, \displaystyle\sum_{i=1}^{k} \frac{\theta_i}{2\pi}\, g_i(x) + \sum_{j=k+1}^{d} \tau_j g_j(x)\Big) \; . \end{array}$$

If M_x is compact, then necessarily $k = d$ and M_x is a torus. The above map then becomes the bijection

$$\mathbb{T}^d \to M_x : \theta \mapsto e\Big(x, \sum_{i=1}^{d} \frac{\theta_i}{2\pi}\, g_i(x)\Big) \; .$$

(b) Next we show that $g_i(x)$ is the gradient of some function $U_i(x)$. For notational convenience, we omit the subscript i and consider a differentiable function g with

$$e(x, g(x)) = e(x, 0) \; , \qquad x \in B \; ,$$

or equivalently,

$$\ell \circ e(x, g(x)) = (x, 0) \; , \qquad x \in B \; .$$

Differentiating this relation gives (with I the d-dimensional identity)

$$A \begin{pmatrix} I \\ g'(x) \end{pmatrix} = \begin{pmatrix} I \\ 0 \end{pmatrix}$$

where A is the Jacobian matrix of $\ell \circ e$ at $(x, g(x))$. We thus have

$$(I \;\; g'(x)^T) A^T J A \begin{pmatrix} I \\ g'(x) \end{pmatrix} = (I \;\; 0) J \begin{pmatrix} I \\ 0 \end{pmatrix} = 0 \; .$$

Since $\ell \circ e$ is a symplectic transformation, we have $A^T J A = J$, and hence the above equation reduces to

$$g'(x)^T - g'(x) = 0 \,.$$

By the Integrability Lemma VI.2.7, there is a function U such that $g(x) = \nabla U(x)$. We may assume $U(x_0) = 0$.

(c) The result of (b) allows us to extend the bijection of (a) to a symplectic transformation. For this, we consider the generating function

$$S(x,\theta) = \sum_{i=1}^{d} \frac{\theta_i}{2\pi}\, U_i(x) \,.$$

With $u(x) = \big(U_1(x), \ldots, U_d(x)\big)$, the mixed second derivative of S is

$$S_{x\theta}(x,\theta) = \frac{1}{2\pi} u_x(x) = \frac{1}{2\pi}\Big(g_1(x), \ldots, g_d(x)\Big) \,,$$

which is invertible because of the linear independence of the g_i. The equations

$$a = \frac{\partial S}{\partial \theta} = \frac{1}{2\pi}\, u(x) \,, \qquad y = \frac{\partial S}{\partial x} = \sum_{i=1}^{d} \frac{\theta_i}{2\pi}\, g_i(x)$$

define a bijective symplectic transformation (for some neighbourhood D of 0, and possibly with a reduced neighbourhood B of x_0)

$$\beta : D \times \mathbb{R}^d \to B \times \mathbb{R}^d : (a,\theta) \mapsto (x,y) = \Big(f(a), \sum_{i=1}^{d} \frac{\theta_i}{2\pi}\, g_i(f(a))\Big)$$

where $x = f(a)$ is the inverse map of $a = \frac{1}{2\pi} u(x)$. We now define

$$\widehat{\psi} = e \circ \beta : D \times \mathbb{R}^d \to \bigcup_{x \in B} M_x \,.$$

By construction, this map is smooth and symplectic, and such that $f_i(a) = x_i = F_i(p,q)$ for $(p,q) = \widehat{\psi}(a,\theta)$. It is surjective by Theorem 1.4. By part (a) of this proof, it becomes injective when the θ_i are taken mod 2π, thus yielding a transformation ψ defined on $D \times \mathbb{T}^d$ with the stated properties. □

X.1.4 Conditionally Periodic Flows

An immediate and important consequence of Theorem 1.6 is the following.

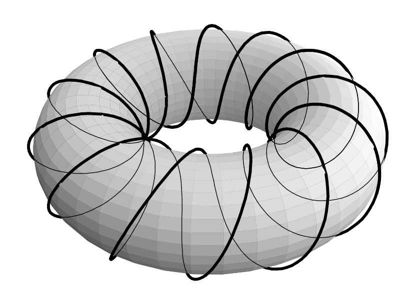
Corollary 1.9. *In the situation of Theorem 1.6, consider the completely integrable system with Hamiltonian $H = F_1$. In the action-angle variables (a,θ), the Hamiltonian equations become*

$$\dot{a}_i = 0, \qquad \dot{\theta}_i = \omega_i(a) \qquad (i = 1, \ldots, d)$$

with $\omega_i(a) = \partial K / \partial a_i(a)$, where $K(a) = H(p,q)$ for $(p,q) = \psi(a,\theta)$.

The flow of a differential system

$$\dot{\theta} = \omega , \qquad \omega = (\omega_i) \in \mathbb{R}^d$$



on the torus $\mathbb{T}^d$ is called *conditionally periodic* with *frequencies* ω_i. The flow is periodic if there exist integers k_i such that for any two frequencies the relation $\omega_i/\omega_j = k_i/k_j$ holds. Otherwise, the flow is called *quasi-periodic*. In particular, the latter occurs when the frequencies are rationally independent, or *non-resonant*: the only integers k_i with $k_1\omega_1 + \ldots + k_d\omega_d = 0$ are $k_1 = \ldots = k_d = 0$. For non-resonant frequencies, it is well known (see Arnold (1989), p. 287) that every trajectory $\{\theta(t) : t \in \mathbb{R}\}$ is dense on the torus $\mathbb{T}^d$ and uniformly distributed.

Example 1.10. We take up again the example of motion in a central field, Example 1.2. For given H and L, we now assume that

$$\{r > 0 \,;\, 2(H - V(r)) - L^2/r^2 > 0\} = [r_0, r_1]$$

is a non-empty interval and the derivatives of $2(H - V(r)) - L^2/r^2$ are non-vanishing at r_0, r_1. By (1.7), the motion from r_0 to r_1 and back again takes a time T and runs through an angle Φ which are given by

$$T = 2\int_{r_0}^{r_1} \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}}\, d\rho\,, \tag{1.12}$$

$$\Phi = 2\int_{r_0}^{r_1} \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}}\, d\rho\,. \tag{1.13}$$

Note that r_0, r_1, T, Φ are functions of H and L. The solution is periodic if Φ is a rational multiple of 2π. This occurs for the Kepler problem, where $\Phi = 2\pi$ and where $T = 2\pi/(-2H)^{3/2}$ (for $H < 0$) depends only on H; see Exercise I.5.

We now construct action-angle variables and compute the frequencies of the system. We begin by constructing the mapping $e(x, y)$ as defined by (1.11) for the variables $x = (x_1, x_2) = (H, L)$ and $y = (y_1, y_2)$ of (1.6). For a given (x, y), we consider $(x, 0)$ and we fix (p, q) with $p = (p_r, p_\varphi)$ and $q = (r, \varphi)$ such that $\ell(p, q) = (x, 0)$, e.g., by choosing $r = r_0$, $\varphi = 0$, $p_r = 0$, $p_\varphi = L$. The mapping $e(x, y)$ is defined by the flow at time $t = 1$ corresponding to the Hamiltonian

$$F_y = y_1 H + y_2 L = y_1\Big(\tfrac{1}{2}(p_r^2 + p_\varphi^2/r^2) + V(r)\Big) + y_2 p_\varphi\,,$$

i.e., by the solution at $t = 1$ of

$$\begin{aligned} \dot{p}_r &= -y_1 \frac{p_\varphi^2}{r^3} - y_1 V'(r)\,, \quad \dot{p}_\varphi = 0 \\ \dot{r} &= y_1 p_r\,, \quad \dot{\varphi} = y_1 \frac{p_\varphi}{r^2} + y_2\,. \end{aligned} \tag{1.14}$$

If we denote the flow of the original system with Hamiltonian $H(p_r, p_\varphi, r, \varphi)$ by φ_t, then we have

$$e(x,y) = \varphi_{y_1}(0, L, r_0, 0) + (0,0,0,y_2)^T$$

with the last component taken modulo 2π. Hence, the values of y satisfying $e(x,y) = e(x,0)$ are

$$y = m_1 g_1(x) + m_2 g_2(x)$$

with integers m_1, m_2 and

$$g_1 = \begin{pmatrix} T \\ -\Phi \end{pmatrix}, \qquad g_2 = \begin{pmatrix} 0 \\ 2\pi \end{pmatrix}.$$

We know from the proof of Theorem 1.6 that g_1 and g_2 are the gradients of functions $U_1(H,L)$ and $U_2(H,L)$, respectively. Clearly, $U_2 = 2\pi L$. The expression for U_1 is less explicit. With the construction of the Integrability Lemma VI.2.7, this function is obtained by quadrature, in a neighbourhood of (H_0, L_0), as

$$U_1(H,L) = \int_0^1 \Big((H - H_0)\, T(H_0 + s(H-H_0), L_0 + s(L-L_0)) - \\ (L-L_0)\, \Phi(H_0 + s(H-H_0), L_0 + s(L-L_0)) \Big) ds\,.$$

(For the Kepler problem, $T = 2\pi/(-2H)^{3/2}$, $\Phi = 0 \bmod 2\pi$, and hence $U_1 = 2\pi/\sqrt{-2H}$.) For the action variables we thus obtain

$$a_1 = \frac{1}{2\pi}\, U_1(H,L)\,, \qquad a_2 = L\,.$$

The angle variables are given by $y = \frac{1}{2\pi}(\theta_1 g_1 + \theta_2 g_2)$, i.e.,

$$\theta_1 = y_1 \frac{2\pi}{T}\,, \qquad \theta_2 = y_2 + y_1 \frac{\Phi}{T}\,. \tag{1.15}$$

Writing the total energy $H = K(a_1, L)$ if a_1 is given by the above formula, we obtain, by differentiation of the identity $2\pi a_1 = U_1(K(a_1,L),L)$,

$$2\pi = \frac{\partial U_1}{\partial H}\frac{\partial K}{\partial a_1}\,, \qquad 0 = \frac{\partial U_1}{\partial H}\frac{\partial K}{\partial a_2} + \frac{\partial U_1}{\partial L}$$

and hence the frequencies

$$\omega_1 = \frac{\partial K}{\partial a_1} = \frac{2\pi}{T}\,, \qquad \omega_2 = \frac{\partial K}{\partial a_2} = \frac{\Phi}{T}\,. \tag{1.16}$$

X.1.5 The Toda Lattice – an Integrable System

> Our method is based on the realization that the Toda lattice belongs to a class of evolution equations which can be studied, and in some cases solved, by utilization of a certain associated eigenvalue problem.
>
> (H. Flaschka 1974)

Classical examples of integrable systems from mechanics include Kepler's problem (Newton 1687/1713, Joh. Bernoulli 1710), the planar motion of a point mass attracted by two fixed centres (Euler 1760), Kepler's problem in a homogeneous force field (Lagrange 1766 solved this as the limit of the previous problem when one centre is at infinity), various spinning tops (Euler 1758b, Lagrange 1788, Kovalevskaya 1889, Goryachev 1899 and Chaplygin 1901), a number of integrable cases of the motion of a rigid body in a fluid, the motion of point vortices in the plane. We refer to Arnold, Kozlov & Neishtadt (1997) and Kozlov (1983) for interesting accounts of these problems and for further references.

Here we consider the celebrated example of the Toda lattice which was the starting point for a huge amount of work on integrable systems in the last few decades, with fascinating relationships to soliton theory in partial differential equations (most notably the Korteweg-de Vries equation) and to eigenvalue algorithms of Numerical Analysis; see Deift (1996) for an account of these developments.

The Toda lattice (or chain) is a system of particles on a line interacting pairwise with exponential forces. Such systems were studied by Toda (1970) as discrete models for nonlinear wave propagation. The motion is determined by the Hamiltonian

$$H(p,q) = \sum_{k=1}^{n} \Big(\frac{1}{2}\, p_k^2 + \exp(q_k - q_{k+1}) \Big) . \tag{1.17}$$

Two types of boundary conditions have found particular attention in the literature:

(i) periodic boundary conditions: $q_{n+1} = q_1$;

(ii) put formally $q_{n+1} = +\infty$, so that the term $\exp(q_n - q_{n+1})$ does not appear.

It was found by Hénon, Flaschka and independently Manakov in 1974 that the periodic Toda system is integrable. Moser (1975) then gave a detailed study of the non-periodic case (ii).

Flaschka (1974) introduced new variables

$$a_k = -\tfrac{1}{2} p_k \, , \quad b_k = \tfrac{1}{2} \exp\big(\tfrac{1}{2}(q_k - q_{k+1}) \big) .$$

(Take $b_n = 0$ in case (ii)). Along a solution $(p(t), q(t))$ of the Toda system, the corresponding functions $(a(t), b(t))$ satisfy the differential equations

$$\dot a_k = 2(b_k^2 - b_{k-1}^2) \, , \quad \dot b_k = b_k (a_{k+1} - a_k)$$

(with $a_{n+1} = a_1$ in case (i), $b_n = 0$ in case (ii)). With the matrices

$$L = \begin{pmatrix} a_1 & b_1 & & & & b_n \\ b_1 & a_2 & b_2 & & 0 & \\ & b_2 & a_3 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & b_{n-2} & a_{n-1} & b_{n-1} \\ b_n & & & & b_{n-1} & a_n \end{pmatrix},$$

$$B = B(L) = \begin{pmatrix} 0 & b_1 & & & & -b_n \\ -b_1 & 0 & b_2 & & 0 & \\ & -b_2 & 0 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & -b_{n-2} & 0 & b_{n-1} \\ b_n & & & & -b_{n-1} & 0 \end{pmatrix},$$

the differential equations can be written in the *Lax pair* form

$$\dot{L} = BL - LB \, . \tag{1.18}$$

This system has an *isospectral flow*, that is, along any solution $L(t)$ of (1.18) the eigenvalues do not depend on t; see Lemma IV.3.4. The eigenvalues $\lambda_1, \ldots, \lambda_n$ of L are therefore first integrals of the Toda system. They are independent and turn out to be in involution, in a neighbourhood of every point where the λ_i are all different; see Exercise 6. Hence, the Toda lattice is a completely integrable system. Its Hamiltonian can be written as

$$H = \sum_{k=1}^{n} \bigl(2a_k^2 + 4b_k^2\bigr) = 2 \,\mathrm{trace}\, L^2 = 2 \sum_{i=1}^{n} \lambda_i^2 \, .$$

We conclude this section with a numerical example for the periodic Toda lattice. We choose $n = 3$ and the initial conditions $p_1 = -1.5$, $p_2 = 1$, $p_3 = 0.5$ and $q_1 = 1$, $q_2 = 2$, $q_3 = -1$. We apply to the system with Hamiltonian (1.17) the symplectic second-order Störmer–Verlet method and the non-symplectic classical fourth-order Runge–Kutta method with two different step sizes. The left pictures of Fig. 1.3 show the numerical approximations to the eigenvalues, and the right pictures the deviations of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ along the numerical solution from their initial values. Clearly, the eigenvalues are not invariants of the numerical schemes. However, Fig. 1.3 illustrates that the eigenvalues along the numerical solution remain close to their correct values over very long time intervals for the symplectic method, whereas they drift off for the non-symplectic method.

An explanation of the long-time near-preservation of the first integrals of completely integrable systems by symplectic methods will be given in the following sections, using backward error analysis and the perturbation theory for integrable Hamiltonian systems.

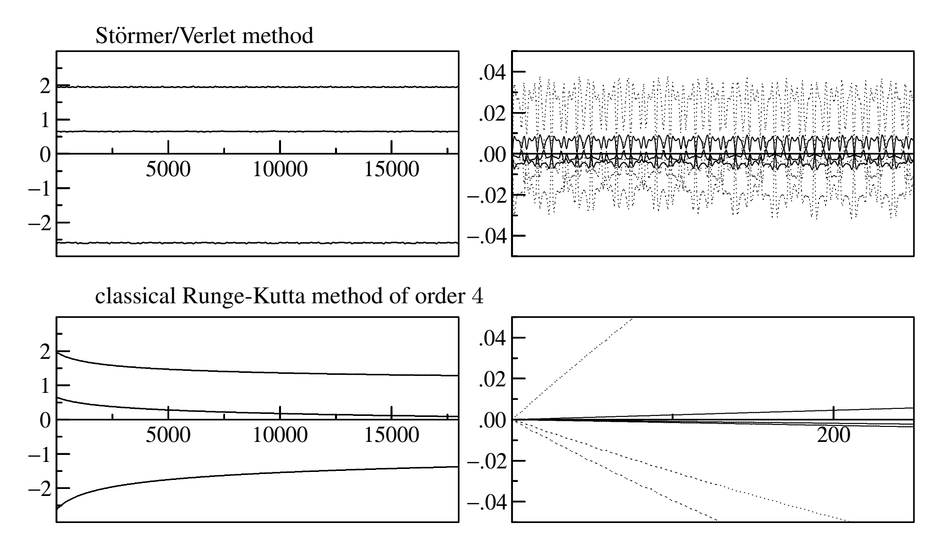


Fig. 1.3. Numerically obtained eigenvalues (left pictures) and errors in the eigenvalues (right pictures) for the step sizes $h = 0.1$ (dotted) and $h = 0.05$ (solid line)

X.2 Transformations in the Perturbation Theory for Integrable Systems

Problème général de la Dynamique. Nous sommes donc conduit à nous proposer le problème suivant: Étudier les équations canoniques

$$\frac{dx_i}{dt} = \frac{d\mathrm{F}}{dy_i}, \qquad \frac{dy_i}{dt} = -\frac{d\mathrm{F}}{dx_i},$$

en supposant que la fonction F peut se développer suivant les puissances d'un paramétre très petit μ de la manière suivante:

$$\mathrm{F} = \mathrm{F}_0 + \mu \mathrm{F}_1 + \mu^2 \mathrm{F}_2 + \ldots ,$$

en supposant de plus que F_0 ne dépend que des x et est indépendant des y; et que $\mathrm{F}_1, \mathrm{F}_2, \ldots$ sont des fonctions périodiques de période 2π par rapport aux y. (H. Poincaré 1892, p. 32f.)

Consider a small perturbation of a completely integrable Hamiltonian. In action-angle variables (a, θ) on $D \times \mathbb{T}^d$ (D an open subset of $\mathbb{R}^d$), this takes the form

$$H(a, \theta) = H_0(a) + \varepsilon H_1(a, \theta) \ , \tag{2.1}$$

where ε is a small parameter. We assume that H_0 and H_1 are real-analytic, and that the perturbation H_1 (which may depend also on ε) is bounded by a constant on a complex neighbourhood of $D \times \mathbb{T}^d$ that is independent of ε. No other restriction shall be imposed on the perturbation.

For the unperturbed system ($\varepsilon = 0$) we have seen that the motion is conditionally periodic on invariant tori $\{a = \text{const.}, \theta \in \mathbb{T}^d\}$. Perturbation theory aims at an understanding of the flow of the perturbed system. The basic tools are symplectic

coordinate transformations which take the system to a form that allows the long-time behaviour (perpetually, or over time scales large compared to ε^{-1}) of solutions of the system (certain solutions, or all solutions with initial values in some ball) to be read off. There are different transformations that provide answers to these problems. The emphasis in this section will be on the construction of suitable transformations, not on the technical but equally important aspects of obtaining estimates for them.

The methods in Poincaré's *Méthodes Nouvelles* form the now classical part of perturbation theory, but the theories of Birkhoff, Siegel, Kolmogorov/Arnold/Moser (KAM) and Nekhoroshev in the 20th century have become "classics" in their own right.

X.2.1 The Basic Scheme of Classical Perturbation Theory

In the spirit of the preceding section, one might search for a symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ close to the identity such that the perturbed Hamiltonian written in the new variables (b, φ) depends only on b, or more modestly, depends only on b up to a remainder term of order $\mathcal{O}(\varepsilon^N)$ with a large $N > 1$, or to begin even more modestly, with $N = 2$. We search for a generating function

$$S(b, \theta) = b \cdot \theta + \varepsilon S_1(b, \theta)$$

where $\cdot$ symbolizes the Euclidean product of vectors in $\mathbb{R}^d$ and S_1 is 2π-periodic in θ. Naively, we require that the symplectic transformation defined by

$$a = \frac{\partial S}{\partial \theta}(b, \theta) \, , \quad \varphi = \frac{\partial S}{\partial b}(b, \theta)$$

be such that the order-ε term in the expansion of the Hamiltonian in the new variables, $K(b, \varphi) = H(a, \theta)$, $K(b, \varphi) = H_0(b) + \varepsilon K_1(b, \varphi) + \ldots$ depends only on b. Since

$$H(a, \theta) = H\Big(b + \varepsilon \frac{\partial S_1}{\partial \theta}(b, \theta), \theta\Big) = H_0(b) + \varepsilon \left\{ \omega(b) \cdot \frac{\partial S_1}{\partial \theta}(b, \theta) + H_1(b, \theta) \right\} + \ldots$$

with the vector of frequencies

$$\omega(b) = \frac{\partial H_0}{\partial b}(b) \, ,$$

the function S_1 must satisfy the partial differential equation

$$\omega(b) \cdot \frac{\partial S_1}{\partial \theta}(b, \theta) + H_1(b, \theta) = \overline{H}_1(b) \tag{2.2}$$

for a function $\overline{H}_1$ that does not depend on θ. Since S_1 is required to be 2π-periodic in θ, the function $\overline{H}_1$ must equal the average of H_1 over the angles:

$$\overline{H}_1(b) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} H_1(b, \theta) \, d\theta \, .$$

Equation (2.2) is the basic equation of Hamiltonian perturbation theory. From the Fourier series of S_1 and H_1,

$$S_1(b,\theta) = \sum_{k\in\mathbb{Z}^d} s_k(b)\, e^{ik\cdot\theta} , \qquad H_1(b,\theta) = \sum_{k\in\mathbb{Z}^d} h_k(b)\, e^{ik\cdot\theta}$$

we obtain a formal solution of (2.2) by comparing Fourier coefficients: $s_0(b)$ is arbitrary and

$$s_k(b) = -\frac{h_k(b)}{ik\cdot\omega(b)} , \qquad k \neq 0 . \tag{2.3}$$

At this point, however, we are struck by the *problem of small denominators*. For any values of the frequencies $\omega_j(b)$, the denominator $k\cdot\omega(b) = k_1\omega_1(b)+\ldots+k_d\omega_d(b)$ becomes arbitrarily small for some $k = (k_1,\ldots,k_d) \in \mathbb{Z}^d$, and even vanishes if the frequencies are rationally dependent.

For a perturbation where only finitely many Fourier coefficients h_k are non-zero, the construction above excludes only a finite number of resonant frequencies (i.e., those with $k\cdot\omega(b) = 0$ for a $k\in\mathbb{Z}^d$ with $h_k \neq 0$) and small neighbourhoods around them. For $\omega(b)$ outside these neighbourhoods and for φ on a complex neighbourhood of $\mathbb{T}^d$, we obtain for the Hamiltonian in the new variables

$$K(b,\varphi) = H_0(b) + \varepsilon\overline{H}_1(b) + \mathcal{O}(\varepsilon^2) .$$

In the general case, we can approximate the perturbation H_1 up to $\mathcal{O}(\varepsilon^2)$ by a trigonometric polynomial. For analytic H_1, the Fourier coefficients h_k decay exponentially with $|k| = \sum_i |k_i|$, and hence the required degree m of the approximating trigonometric polynomial grows logarithmically with ε, i.e., $m \sim |\log\varepsilon|$.

As $\varepsilon \to 0$, the remainder term is under control only for those frequencies $\omega = \omega(b)$ for which the exponentially decaying Fourier coefficients h_k of the perturbation decay faster than the denominators $ik\cdot\omega$ with growing $|k|$. This is certainly the case for frequencies satisfying *Siegel's diophantine condition* (or *strong non-resonance condition*, as it is sometimes called)

$$|k\cdot\omega| \geq \gamma\,|k|^{-\nu} , \qquad k\in\mathbb{Z}^d, k\neq 0 \tag{2.4}$$

for some positive constants γ, ν. (Here again, $|k| = \sum_i |k_i|$). If $\nu > d-1$, the set of frequencies in a fixed ball that do *not* satisfy (2.4) has Lebesgue measure bounded by $Const\cdot\gamma$ (Exercise 5). Therefore, almost all frequencies satisfy (2.4) for some $\gamma > 0$. However, for any γ and ν, the complementary set is open and dense in $\mathbb{R}^d$.

X.2.2 Lindstedt–Poincaré Series

> ... pour que la méthode de M. Lindstedt soit applicable, soit sous sa forme primitive, soit sous celle que je lui ai ensuite donnée, il faut qu'en première approximation les moyens mouvements ne soient liés par aucune relation linéaire à coefficients entiers; ...
>
> Il semble donc permis de conclure que les séries (...) ne convergent pas. Toutefois le raisonnement qui précède ne suffit pas pour établir ce point avec une rigueur complète. (H. Poincaré 1893, pp. vi, 103.)



Fig. 2.1. Henri Poincaré (left), born: 29 April 1854 in Nancy (France), died: 17 July 1912 in Paris; Anders Lindstedt (right), born: 27 June 1854 in Sundborn (Sweden), died: 1939. Reproduced with permission of Bibl. Math. Univ. Genève

The above construction is extended without any additional difficulty to arbitrary finite order in ε. The generating function is now sought in the form

$$S(b,\theta) = b\cdot\theta + \varepsilon S_1(b,\theta) + \varepsilon^2 S_2(b,\theta) + \ldots + \varepsilon^{N-1} S_{N-1}(b,\theta) \tag{2.5}$$

and, as before, the requirement that the first N terms in the ε-expansion of the Hamiltonian in the new variables be independent of the angles, leads via a Taylor expansion of the Hamiltonian to equations of the form (2.2) for $S_1,\ldots,S_{N-1}$:

$$\omega(b)\cdot\frac{\partial S_j}{\partial\theta} + K_j(b,\theta) = \overline{K}_j(b) \tag{2.6}$$

where $K_1 = H_1$,

$$K_2 = \frac{1}{2}\frac{\partial^2 H_0}{\partial a^2}\left(\frac{\partial S_1}{\partial\theta},\frac{\partial S_1}{\partial\theta}\right) + \frac{\partial H_1}{\partial a}\cdot\frac{\partial S_1}{\partial\theta}\ ,$$

and in general, K_j is a sum of terms

$$\frac{1}{i!}\frac{\partial^i H_{k_0}}{\partial a^i}\left(\frac{\partial S_{k_1}}{\partial\theta},\ldots,\frac{\partial S_{k_i}}{\partial\theta}\right) \qquad \text{with} \quad k_0+k_1+\ldots+k_i = j\ .$$

The function $\overline{K}_j$ denotes again the angular average of K_j. These equations can be formally solved in the case of rationally independent frequencies. The Hamiltonian in the new variables is then

$$K(b,\varphi) = H_0(b)+\varepsilon\overline{K}_1(b)+\varepsilon^2\overline{K}_2(b)+\ldots+\varepsilon^{N-1}\overline{K}_{N-1}(b)+\varepsilon^N R_N(b,\theta)\,. \quad (2.7)$$

The possible convergence of the series for $N \to \infty$ is a delicate issue that was not resolved conclusively by Poincaré (1893) in his chapter on "Divergence des séries de M. Lindstedt". If for some b^*, the series (2.5) together with its partial derivatives converged as $N \to \infty$, then $\{b = b^*, \varphi \in \mathbb{T}^d\}$ would be an invariant torus of the perturbed Hamiltonian system. However, it was not until Kolmogorov (1954) that the existence of invariant tori – for diophantine frequencies – was found, using a different construction. A direct proof of the convergence of the series of classical perturbation theory for diophantine frequencies was obtained only in 1988 by Eliasson (published in 1996); also see Giorgilli & Locatelli (1997) and references therein.

Nevertheless, already the truncated series (2.5) leads in a rather simple way to strong conclusions about the flow over long time scales when it is combined with the idea of approximating the Hamiltonian by a trigonometric polynomial: the "ultra-violet cut-off", an idea briefly addressed by Poincaré (1893), p. 98f., and taken to its full bearing by Arnold (1963) in his proof of the KAM theorem. We formulate a lemma for a fixed truncation index N. Here, $\omega_{\varepsilon,N}(b)$ denotes the derivative of the truncated series (2.7) with respect to b.

Lemma 2.1. *Suppose that $\omega(b^*)$ satisfies the diophantine condition (2.4). For any fixed $N \geq 2$, there are positive constants ε_0, c, C such that the following holds for $\varepsilon \leq \varepsilon_0$: there exists a real-analytic symplectic change of coordinates $(a,\theta) \mapsto (b,\varphi)$ such that every solution $(b(t),\varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \leq c|\log\varepsilon|^{-\nu-1}$, satisfies*

$$\|b(t) - b(0)\| \leq C\,t\,\varepsilon^N \qquad \text{for} \quad t \leq \varepsilon^{-N+1}\,,$$

$$\|\varphi(t) - \omega_{\varepsilon,N}(b(0))t - \varphi(0)\| \leq C\,(t^2 + t|\log\varepsilon|^{\nu+1})\,\varepsilon^N \quad \text{for} \;\; t^2 \leq \varepsilon^{-N+1}.$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$-close to the identity: $\|(a,\theta) - (b,\varphi)\| \leq C\varepsilon$ holds for (a,θ) and (b,φ) related by the above coordinate transform, for $\|b-b^*\| \leq c|\log\varepsilon|^{-\nu-1}$ and for φ in an ε-independent complex neighbourhood of $\mathbb{T}^d$.

The constants ε_0, c, C depend on N, d, γ, ν and on bounds of H_0 and H_1 on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$.

Proof. Using the relations (2.3) and their analogues for (2.6), it is a straightforward but somewhat tedious exercise to show that at the given particular b^*, the functions $K_j(b^*,\cdot)$, $S_j(b^*,\cdot)$ are all analytic on the same complex neighbourhood of $\mathbb{T}^d$, and that the remainder term is bounded by

$$|R_N(b^*,\theta)| \leq C = C(N,d,\gamma,\nu)$$

for all θ in a complex neighbourhood of $\mathbb{T}^d$ which is independent of ε. Here, C depends in addition on the bound of H_1 on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$, or what amounts to the same by Cauchy's estimates, on bounds of the exponential decay of the Fourier coefficients h_k of H_1. (In case of doubt, see also Sect. X.4 for explicit estimates.)

Assume first that $H_1(b,\theta)$ is a trigonometric polynomial in θ of degree m. Then K_j, S_j are trigonometric polynomials of degree jm. Since $|k\cdot\omega(b)| \geq |k\cdot\omega(b^*)| - |k|(\max\|\omega'\|)\|b-b^*\|$, there is a $\delta > 0$ such that

$$|k\cdot\omega(b)| \geq \tfrac{1}{2}\gamma\,|k|^{-\nu} \quad \text{for} \quad \|b-b^*\| \leq \delta,\;\; |k| \leq Nm.$$

This number δ is proportional to $\gamma(Nm)^{-\nu-1}$. Consequently, since the construction involves only the trigonometric polynomials K_j, S_j of degree up to Nm, the above estimate for the remainder term R_N holds also for $\|b-b^*\| \leq \delta$. To approximate a general analytic H_1 by trigonometric polynomials up to $\mathcal{O}(\varepsilon^N)$, we must choose the degree m proportional to $|\log\varepsilon^N|$. With the choice $\delta = c\,(N^2|\log\varepsilon|)^{-\nu-1}$, for a sufficiently small $c > 0$ independent of ε (and N), the above bound for the remainder $R_N(b,\theta)$ is then valid for b in the complex ball $\|b-b^*\| \leq 2\delta$ and for φ in a complex neighbourhood of $\mathbb{T}^d$ (which depends only on N). By Cauchy's estimates, this implies

$$\left\|\frac{\partial R_N}{\partial\theta}(b,\theta)\right\| \leq C\;, \qquad \left\|\frac{\partial R_N}{\partial b}(b,\theta)\right\| \leq \frac{C}{\delta}$$

for $\|b-b^*\| \leq \delta$ and $\theta \in \mathbb{T}^d$. Hence, as long as $\|b(t)-b^*\| \leq \delta$, the Hamiltonian differential equations are of the form

$$\dot b = -\frac{\partial K}{\partial\varphi} = -\varepsilon^N\,\frac{\partial R_N}{\partial\theta}\,\frac{\partial\theta}{\partial\varphi} = \mathcal{O}(\varepsilon^N)\;, \quad \dot\varphi = \frac{\partial K}{\partial b} = \omega_{\varepsilon,N}(b) + \mathcal{O}(\varepsilon^N/\delta)\;.$$

This implies the result. □

Hence, the tori $\{b = b(0),\ \varphi \in \mathbb{T}^d\}$ are nearly invariant over a time scale ε^{-N+1}, and the flow is close to a quasiperiodic flow over times bounded by the square root of ε^{-N+1}. Lemma 2.1 is just a preliminary to more substantial results (which hold under appropriate additional conditions): invariant tori carrying a quasiperiodic flow with diophantine frequencies persist under small Hamiltonian perturbations (Kolmogorov 1954); every solution of the perturbed system remains close, within a positive power of ε, to some torus over times that are exponentially long in a negative power of ε (Nekhoroshev 1977); solutions starting close to an invariant torus with diophantine frequencies stay within twice the initial distance over time intervals that are exponentially long in a negative power of the distance (Perry & Wiggins 1994) or even exponentially long in the exponential of the inverse of the distance (Morbidelli & Giorgilli 1995).

The symplectic transformations of this subsection were constructed using the mixed-variable generating function $S(b,\theta)$. As was pointed out for example by Benettin, Galgani & Giorgilli (1985), rigorous estimates for the remainder terms are often obtained in a simpler way using the *Lie method*, which involves constructing the near-identity symplectic transformation as the time-ε flow of some auxiliary Hamiltonian system with a suitably defined Hamiltonian $\chi(b,\varphi)$. As before, the condition that the Hamiltonian $H(a,\theta) = K(b,\varphi)$ should depend on φ only in higher-order terms, leads to equations of the form (2.2), now for χ instead of S_1. We will use such a construction in the following subsection.

X.2.3 Kolmogorov's Iteration

> It is easy to grasp the meaning of Theorem 1 for mechanics. It indicates that an s-parametric family of conditionally periodic motions [...] cannot, under conditions (3) and (4) [here: (2.4) and (2.9)], disappear as a result of a small change in the Hamilton function H.
>
> In this note we confine ourselves to the construction of the transformation. (A.N. Kolmogorov 1954)

For the completely integrable Hamiltonian $H_0(a)$, the phase space is foliated into invariant tori parametrized by a. We now fix one such torus $\{a = a^*,\ \theta \in \mathbb{T}^d\}$ with strongly diophantine frequencies $\omega = \omega(a^*)$. Without loss of generality, we may assume $a^* = 0$. This particular torus is invariant under the flow of every Hamiltonian $H(a, \theta)$ for which the linear terms in the Taylor expansion with respect to a at 0 are independent of θ:

$$H(a, \theta) = c + \omega \cdot a + \tfrac{1}{2} a^T M(a, \theta) a \tag{2.8}$$

with $c \in \mathbb{R}$, $\omega \in \mathbb{R}^d$, and a real symmetric $d \times d$-matrix $M(a, \theta)$ analytic in its arguments. Since the Hamiltonian equations are of the form

$$\dot{a} = \mathcal{O}(\|a\|^2) , \quad \dot{\theta} = \omega + \mathcal{O}(\|a\|) ,$$

the torus $\{a = 0,\ \theta \in \mathbb{T}^d\}$ is invariant and the flow on it is quasi-periodic with frequencies ω.

Consider now an analytic perturbation of such a Hamiltonian: $H(a, \theta) + \varepsilon G(a, \theta)$ with a small ε. Kolmogorov (1954) found a near-identity symplectic transformation $(a, \theta) \mapsto (\widetilde{a}, \widetilde{\theta})$, constructed by an iterative procedure, such that the perturbed Hamiltonian in the new variables is again of the form (2.8) with the same ω, and hence has the invariant torus $\{\widetilde{a} = 0,\ \widetilde{\theta} \in \mathbb{T}^d\}$ carrying a quasi-periodic flow with the frequencies of the unperturbed system. This holds under the conditions that ω satisfies the diophantine condition (2.4), and that the angular average

$$\overline{M}_0 := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} M(0, \theta)\, d\theta \quad \text{is an invertible matrix.} \tag{2.9}$$

Here we describe the iterative construction of this symplectic transformation. The proof of convergence of the iteration will be given in Sect. X.5.

We construct a symplectic transformation $(a, \theta) \mapsto (b, \varphi)$ as the time-ε flow of an auxiliary Hamiltonian of the form

$$\chi(b, \varphi) = \xi \cdot \varphi + \chi_0(\varphi) + \sum_{i=1}^{d} b_i \chi_i(\varphi) , \tag{2.10}$$

where $\xi \in \mathbb{R}^d$ is a constant vector, and $\chi_0, \chi_1, \ldots, \chi_d$ are 2π-periodic functions. (Quadratic and higher-order terms in b play no role in the construction and are therefore omitted right at the outset.) The old and new coordinates are then related by

$$a = b + \varepsilon \frac{\partial \chi}{\partial \varphi}(b, \varphi) + \mathcal{O}(\varepsilon^2) , \quad \theta = \varphi - \varepsilon \frac{\partial \chi}{\partial b}(b, \varphi) + \mathcal{O}(\varepsilon^2) .$$

We insert this into

$$H(a,\theta)+\varepsilon G(a,\theta) = c+\omega\cdot b+\tfrac{1}{2}b^T M(b,\varphi)b$$
$$+\,\varepsilon\left\{\omega\cdot\frac{\partial\chi}{\partial\varphi}(b,\varphi)+b^T M(b,\varphi)\frac{\partial\chi}{\partial\varphi}(b,\varphi)+G(b,\varphi)\right\}+\mathcal{O}(\varepsilon\|b\|^2)+\mathcal{O}(\varepsilon^2)\,.$$

We now require that the term in curly brackets be $Const+\mathcal{O}(\|b\|^2)$. Writing down the Taylor expansion

$$G(b,\varphi)=G_0(\varphi)+\sum_{i=1}^{d} b_i G_i(\varphi)+b^T Q(b,\varphi)b \tag{2.11}$$

and inserting the above ansatz for χ, this condition becomes

$$\omega\cdot\frac{\partial\chi_0}{\partial\varphi}(\varphi)+\sum_{i=1}^{d} b_i\left(\omega\cdot\frac{\partial\chi_i}{\partial\varphi}(\varphi)+u_i(\varphi)+v_i(\varphi)\right)$$
$$+\,G_0(\varphi)+\sum_{i=1}^{d} b_i G_i(\varphi)\,=\,Const.,$$

where $u=(u_1,\ldots,u_d)^T$ and $v=(v_1,\ldots,v_d)^T$ are defined by

$$u(\varphi)=M(0,\varphi)\xi\,, \tag{2.12}$$
$$v(\varphi)=M(0,\varphi)\,\frac{\partial\chi_0}{\partial\varphi}(\varphi)\,. \tag{2.13}$$

The condition is fulfilled if

$$\omega\cdot\frac{\partial\chi_0}{\partial\varphi}(\varphi)+G_0(\varphi)=\overline{G}_0 \tag{2.14}$$
$$\omega\cdot\frac{\partial\chi_i}{\partial\varphi}(\varphi)+u_i(\varphi)+v_i(\varphi)+G_i(\varphi)=\overline{u}_i+\overline{v}_i+\overline{G}_i \tag{2.15}$$
$$\overline{u}_i+\overline{v}_i+\overline{G}_i=0 \qquad (i=1,\ldots,d). \tag{2.16}$$

Here the bars again denote angular averages. Note that equations (2.14), (2.15) are of the form (2.2). Equation (2.14) determines χ_0 and hence $v=(v_1,\ldots,v_d)^T$ by (2.13). Equations (2.16) then give $\overline{u}=(\overline{u}_1,\ldots,\overline{u}_d)^T$. By (2.12), we need

$$\overline{u}=\overline{M}_0\xi\,,$$

which determines ξ uniquely because $\overline{M}_0$ is assumed to be invertible. Equation (2.12) then yields $u=(u_1,\ldots,u_d)^T$. Finally, (2.15) determines $\chi_1,\ldots,\chi_d$, and the construction of $\chi(b,\varphi)$ is complete. In the new variables (b,φ), the perturbed Hamiltonian then takes the form

$$H(a,\theta)+\varepsilon G(a,\theta)=\widehat{c}+\omega\cdot b+\tfrac{1}{2}b^T\widehat{M}(b,\varphi)b+\varepsilon^2\widehat{G}(b,\varphi) \tag{2.17}$$

with unchanged frequencies ω and with $\widehat{M}(b,\varphi) = M(b,\varphi) + \mathcal{O}(\varepsilon)$. The perturbation to the form (2.8) is thus reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. The iteration of this procedure turns out to be convergent, see Sect. X.5. This finally yields a symplectic change of coordinates that transforms the perturbed Hamiltonian to the form (2.8). The perturbed system thus has an invariant torus carrying a quasi-periodic flow with frequencies ω – a KAM torus, as it is named after Kolmogorov, Arnold and Moser.

X.2.4 Birkhoff Normalization Near an Invariant Torus

KAM tori are very sticky.
(A.D. Perry & S. Wiggins 1994)

In this subsection we describe a transformation studied by Pöschel (1993) and Perry & Wiggins (1994) for systems with Hamiltonian in the Kolmogorov form (2.8) in a neighbourhood of the invariant torus $\{a = 0,\ \theta \in \mathbb{T}^d\}$. This transformation is an analogue of a transformation of Birkhoff (1927) for Hamiltonian systems near an elliptic stationary point.

The symplectic change of coordinates $(a,\theta) \mapsto (b,\varphi)$ considered here transforms a Hamiltonian (2.8) with diophantine frequencies ω to the form $H(a,\theta) = K_N(b) + \mathcal{O}(\|b\|^N)$ for arbitrary N, or more precisely, the Hamiltonian in the new variables, $H_N(b,\varphi) = H(a,\theta)$, is of the form

$$H_N(b,\varphi) = \omega \cdot b + Z_N(b) + R_N(b,\varphi) \tag{2.18}$$

with $Z_N(b) = \mathcal{O}(\|b\|^2)$ and $R_N(b,\varphi) = \mathcal{O}(\|b\|^N)$. (We have taken the irrelevant constant term in (2.8) $c = 0$.) The equations of motion then take the form

$$\dot b = \mathcal{O}(\|b\|^N)\ , \qquad \dot\varphi = \omega + \mathcal{O}(\|b\|)\ .$$

Therefore, in these variables $\{b = 0, \varphi \in \mathbb{T}^d\}$ is an invariant torus, and for sufficiently small r,

$$\|b(0)\| \le r \qquad \text{implies} \qquad \|b(t)\| \le 2r \ \text{ for } \ t \le C_N\, r^{-N+1}\ .$$

A judicious choice of N even yields time intervals that are exponentially long in a negative power of r on which solutions starting at a distance r stay within twice the initial distance (Perry & Wiggins 1994). Motion away from the torus can thus be only very slow.

The normal form (2.18) is constructed iteratively. Each iteration step is very similar to the procedure in Sect. X.2.1, where now the distance to the torus plays the role of the small parameter. Consider a Hamiltonian

$$H(a,\theta) = \omega \cdot a + Z(a) + R(a,\theta)$$

where $Z(a) = \mathcal{O}(\|a\|^2)$ and $R(a,\theta) = \mathcal{O}(\|a\|^k)$ for some $k \ge 2$ in a complex neighbourhood of $\{0\} \times \mathbb{T}^d$. We construct a symplectic change of coordinates $(a,\theta) \mapsto (b,\varphi)$ via a generating function $b \cdot \theta + S(b,\theta)$ as

$$a = b + \frac{\partial S}{\partial \theta}(b,\theta)\ , \quad \varphi = \theta + \frac{\partial S}{\partial b}(b,\theta)\ .$$

We expand (omitting the arguments (b,θ) in $\partial S/\partial\theta$ and $\partial H/\partial a$)

$$\begin{aligned} H\Big(b + \frac{\partial S}{\partial \theta},\theta\Big) &= H(b,\theta) + \frac{\partial H}{\partial a}\cdot\frac{\partial S}{\partial \theta} + Q(b,\theta) \\ &= \omega\cdot b + Z(b) + \Big\{ R(b,\theta) + \frac{\partial H}{\partial a}\cdot\frac{\partial S}{\partial \theta} \Big\} + Q(b,\theta)\ , \end{aligned}$$

where $|Q(b,\theta)| \le \text{Const.}\,\|\partial S/\partial\theta\|^2$. Since $\partial H/\partial b = \omega + \mathcal{O}(\|b\|)$, we can make the expression in curly brackets independent of θ up to $\mathcal{O}(\|b\|^{k+1})$ by determining S from the equation of the form (2.2):

$$\omega\cdot\frac{\partial S}{\partial \theta}(b,\theta) + R(b,\theta) = \overline{R}(b)\ .$$

For diophantine frequencies ω, we obtain $S(b,\theta) = \mathcal{O}(\|b\|^k)$ on a (reduced) complex neighbourhood of $\{0\}\times\mathbb{T}^d$ from the corresponding estimate for $R(b,\theta)$. It follows that the above symplectic transformation with generating function $b\cdot\theta + S(b,\theta)$ is well-defined for small $\|b\|$, and the Hamiltonian in the new variables, $\widehat{H}(b,\varphi) = H(a,\theta)$, becomes

$$\widehat{H}(b,\varphi) = \omega\cdot b + \widehat{Z}(b) + \widehat{R}(b,\varphi)$$

with $\widehat{Z}(b) = Z(b) + \overline{R}(b)$ and

$$\widehat{R}(b,\varphi) = \Big(\frac{\partial H}{\partial a}(b,\theta) - \omega\Big)\cdot\frac{\partial S}{\partial \theta}(b,\theta) + Q(b,\theta) = \mathcal{O}(\|b\|^{k+1})\ ,$$

so that the order in b of the remainder term is augmented by 1. The procedure can be iterated, but unlike the iteration of the preceding subsection, this iteration is in general divergent. Nevertheless, a suitable finite termination yields remainder terms that are exponentially small in a positive power of r for $\|b\| \le r$, by arguments similar to those of Sect. X.4.

X.3 Linear Error Growth and Near-Preservation of First Integrals

In the remaining part of this chapter we study the long-time behaviour of symplectic discretizations of integrable and near-integrable Hamiltonian systems. While here we will be concerned with general symplectic methods, it should be noted that some integrable problems admit integrable discretizations; see Suris (2003).

In this section we are concerned with the error growth of symplectic numerical methods and their approximate preservation of first integrals. A preliminary analysis of linear error growth for the Kepler problem was first given by Calvo & Sanz-Serna

(1993). Using backward error analysis and KAM theory, Calvo & Hairer (1995a) then showed linear error growth of symplectic methods applied to integrable systems when the frequencies at the initial value satisfy a diophantine condition (2.4). Here we give such a result under milder conditions on the initial values, combining backward error analysis and Lemma 2.1. We derive also a first result on the long-time near-preservation of all first integrals, which will be extended to exponentially long times in Sections X.4.3 and X.5.2 (under stronger assumptions on the starting values), and perpetually in Sect. X.6 (only for a Cantor set of step sizes).

Figure 3.1 illustrates the linear error growth of the symplectic Störmer–Verlet method, as opposed to the quadratic error growth for the classical fourth-order Runge–Kutta method, on the example of the Toda lattice. The same number of function evaluations was used for both methods.

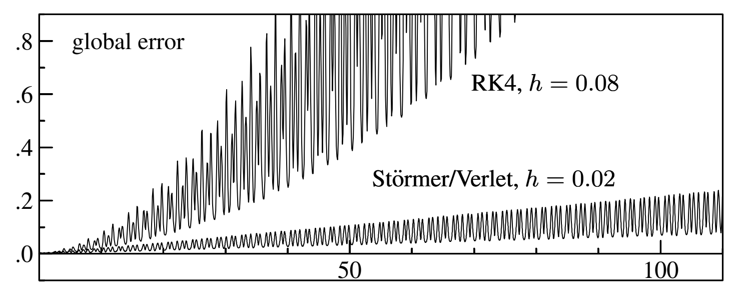


Fig. 3.1. Euclidean norm of the global error for the Störmer–Verlet scheme (step size $h = 0.02$) and the classical Runge–Kutta method of order 4 (step size $h = 0.08$) applied to the Toda lattice with $n = 3$ and initial values as in Fig. 1.3

We consider a completely integrable Hamiltonian system (usually not given in action-angle variables)

$$\dot p = -\frac{\partial H}{\partial q}(p,q) , \quad \dot q = \frac{\partial H}{\partial p}(p,q) \tag{3.1}$$

and apply to it a symplectic numerical method with step size h, yielding a numerical solution sequence (p_n, q_n). We assume that the Hamiltonian is real-analytic and that the conditions of the Arnold–Liouville theorem, Theorem 1.6, are fulfilled. Consider the symplectic transformation $(p, q) = \psi(a, \theta)$ to action-angle variables. We denote the inverse transformation as

$$(a, \theta) = \big(I(p, q), \Theta(p, q)\big) . \tag{3.2}$$

We recall that the components $I_1, \dots, I_d$ of $I = (I_i)$ are first integrals of the system: $I(p(t), q(t)) = I(p_0, q_0)$ for all t. In the action-angle variables, the Hamiltonian is $\mathcal{H}(a) = H(p, q)$, and we denote the frequencies

$$\omega(a) = \frac{\partial \mathcal{H}}{\partial a}(a) . \tag{3.3}$$

We consider this in a neighbourhood of some $a^* \in \mathbb{R}^d$.

Theorem 3.1. *Consider applying a symplectic numerical integrator of order p to the completely integrable Hamiltonian system (3.1). Suppose that $\omega(a^*)$ satisfies the diophantine condition (2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \le h_0$: every numerical solution starting with $\|I(p_0,q_0) - a^*\| \le c|\log h|^{-\nu-1}$ satisfies*

$$\begin{aligned} \|(p_n,q_n) - (p(t),q(t))\| &\le C\, t\, h^p \\ \|I(p_n,q_n) - I(p_0,q_0)\| &\le C\, h^p \end{aligned} \qquad \textit{for } t = nh \le h^{-p} .$$

The constants h_0, c, C depend on d, γ, ν, on bounds of the real-analytic Hamiltonian H on a complex neighbourhood of the torus $\{(p,q)\,;\, I(p,q) = a^*\}$, and on the numerical method.

Proof. (a) In the action-angle variables (a,θ), the exact flow is given as

$$a(t) = a(0)\ , \quad \theta(t) = \omega(a(0))\, t + \theta(0)\ . \tag{3.4}$$

By Theorem IX.3.1 (and Theorem IX.1.2), the truncated modified equation of the numerical method is Hamiltonian with[1]

$$\widetilde{H}(p,q) = H(p,q) + h^p H_{p+1}(p,q) + \ldots + h^r H_{r+1}(p,q)\ .$$

We choose $r = 2p$, and we denote by $(\widetilde{p}(t), \widetilde{q}(t))$ the solution of the modified equations with initial values (p_0,q_0). In the variables (a,θ), the modified Hamiltonian becomes $\widetilde{H}(p,q) = \widetilde{\mathcal{H}}(a,\theta)$ with

$$\widetilde{\mathcal{H}}(a,\theta) = \mathcal{H}(a) + \varepsilon\, \mathcal{G}_h(a,\theta)\ , \tag{3.5}$$

where $\varepsilon = h^p$ and the perturbation function $\mathcal{G}_h$ is bounded independently of h on a complex neighbourhood of $\{a^*\} \times \mathbb{T}^d$. By Lemma 2.1 with $\varepsilon = h^p$ and $N \ge 3$, there is a symplectic change of coordinates $\mathcal{O}(h^p)$-close to the identity, such that the solution of the modified equation in the new variables (b,φ) is of the form

$$\begin{aligned} \widetilde{b}(t) &= \widetilde{b}(0) + \mathcal{O}(th^{pN})\ , \\ \widetilde{\varphi}(t) &= \omega_h(\widetilde{b}(0))\, t + \widetilde{\varphi}(0) + \mathcal{O}(th^{pN-1} + t^2 h^{pN}) \end{aligned} \qquad \text{for } t \le h^{-p}\ , \tag{3.6}$$

with $\omega_h(b) = \omega(b) + \mathcal{O}(h^p)$. The constants symbolized by the $\mathcal{O}$-notation are independent of h, of $t \le h^{-p}$ and of $(\widetilde{b}(0), \widetilde{\varphi}(0))$ with $|\widetilde{b}(0) - a^*| \le c|\log h|^{-\nu-1}$. Since the transformation between the variables (a,θ) and (b,φ) is $\mathcal{O}(h^p)$ close to the identity, it follows that the flow of the modified equations in the variables (a,θ) satisfies

[1] We always assume, without further mention, that the modified Hamiltonian is well-defined on the same open set D as the original Hamiltonian. This is true for arbitrary symplectic methods if D is simply connected; on general domains it is satisfied for (partitioned) Runge–Kutta methods and for splitting methods; see Sections IX.3 and IX.4.

$$\begin{aligned} \widetilde{a}(t) &= \widetilde{a}(0) + \mathcal{O}(h^p)\,, \\ \widetilde{\theta}(t) &= \omega(\widetilde{a}(0))\,t + \widetilde{\theta}(0) + te_h + \mathcal{O}(h^p) \end{aligned} \qquad \text{for } 1 \le t \le h^{-p}\,,$$

where $e_h = \omega_h(\widetilde{b}(0)) - \omega(\widetilde{a}(0)) = \mathcal{O}(h^p)$ yields the dominant contribution to the error. By comparison with (3.4) and since $\widetilde{a}(t) = I(\widetilde{p}(t), \widetilde{q}(t))$, the difference between the exact solution and the solution of the modified equation therefore satisfies

$$\begin{aligned} (\widetilde{p}(t), \widetilde{q}(t)) - (p(t), q(t)) &= \mathcal{O}(th^p) \\ I(\widetilde{p}(t), \widetilde{q}(t)) - I(p_0, q_0) &= \mathcal{O}(h^p) \end{aligned} \qquad \text{for } 1 \le t \le h^{-p}.$$

The same bounds for $t \le 1$ follow by standard error estimates.

(b) It remains to bound the difference between the solution of the modified equation and the numerical solution. By construction of the modified equation with $r = 2p$ and by comparison with (3.6), one step of the method is of the form

$$b_{n+1} = b_n + \mathcal{O}(h^{r+1})\,, \quad \varphi_{n+1} = \omega_h(b_n)\,h + \varphi_n + \mathcal{O}(h^{r+1}).$$

It follows that for $t = nh$,

$$b_n = \widetilde{b}(t) + \mathcal{O}(th^r)\,, \quad \varphi_n = \widetilde{\varphi}(t) + \mathcal{O}(t^2 h^r)\,.$$

For $t \le h^{-p}$ and $r = 2p$, we have $th^r \le h^p$. Hence the difference between the numerical solution and the solution of the modified equations in the original variables (p, q) is bounded by

$$\begin{aligned} (p_n, q_n) - (\widetilde{p}(t), \widetilde{q}(t)) &= \mathcal{O}(th^p) \\ I(p_n, q_n) - I(\widetilde{p}(t), \widetilde{q}(t)) &= \mathcal{O}(h^p) \end{aligned} \qquad \text{for } t = nh \le h^{-p}\,.$$

Together with the bound of part (a) this gives the result. □

Remark 3.2. The linear error growth holds also when the symplectic method is applied to a perturbed integrable system with a perturbation parameter ε bounded by a positive power of the step size: $\varepsilon \le K\,h^\alpha$ for some $\alpha > 0$. The proof of this generalization is the same as above, except that possibly a larger N is required in using Lemma 2.1.

Example 3.3 (Linear Error Growth for the Kepler Problem). From Example 1.10 we know that for the Kepler problem the frequencies (1.16) do not satisfy the diophantine condition (2.4). Nevertheless we observed a linear error growth for symplectic methods in the experiments of Fig. I.2.3 (see also Table I.2.1). This can be explained as follows: in action-angle variables the Hamiltonian of the Kepler problem is $\mathcal{H}(a_1, a_2)$, where $a_2 = L$ is the angular momentum. Since the angular momentum is a quadratic invariant that is exactly conserved by symplectic integrators such as symplectic partitioned Runge–Kutta methods, the modified Hamiltonian

$$\widetilde{\mathcal{H}}(a, \theta) = \mathcal{H}(a_1, a_2) + \varepsilon\,\mathcal{G}_h(a_1, a_2, \theta_1)$$

does not depend on the angle variable θ_2 (see Corollary IX.5.3). As in the proof of Lemma 2.1 we average out the angle θ_1 up to a certain power of ε. Since we are concerned here with one degree of freedom, the diophantine condition is trivially satisfied, and we can conclude as in Theorem 3.1.

X.4 Near-Invariant Tori on Exponentially Long Times

We refine the results for the classical perturbation series of Sect. X.2.2 to yield locally integrable behaviour, up to exponentially small deviations, over time intervals that are exponentially long in a power of the small perturbation parameter. We then combine this result with backward error analysis to show the near-preservation of invariant tori over exponentially long times in a negative power of the step size for symplectic integrators. We begin with the necessary technical estimates.

X.4.1 Estimates of Perturbation Series

We will estimate the coefficients of the perturbation series (2.5), which requires a bound for the solution of (2.6). We use the following notation: for $\rho > 0$ and with $\|\cdot\|$ the maximum norm on $\mathbb{R}^d$,

$$U_\rho = \{\theta \in \mathbb{T}^d + i\mathbb{R}^d \; ; \; \|\mathrm{Im}\,\theta\| < \rho\}$$

denotes the complex extension of the d-dimensional torus $\mathbb{T}^d$ of width ρ. For a bounded analytic function F on U_ρ, we write

$$\|F\|_\rho = \sup_{\theta \in U_\rho} |F(\theta)| \; , \qquad \Big\| \frac{\partial F}{\partial \theta} \Big\|_\rho = \sum_{j=1}^{d} \Big\| \frac{\partial F}{\partial \theta_j} \Big\|_\rho \; .$$

Following Arnold (1963), we prove the following bounds for the solution of the basic partial differential equation (2.2) .

Lemma 4.1. *Suppose $\omega \in \mathbb{R}^d$ satisfies the diophantine condition (2.4). Let G be a bounded real-analytic function on U_ρ, and let $\overline{G}$ denote the average of G over $\mathbb{T}^d$. Then, the equation*

$$\omega \cdot \frac{\partial F}{\partial \theta} + G = \overline{G}$$

has a unique real-analytic solution F on U_ρ with zero average $\overline{F} = 0$. For every positive $\delta < \min(\rho, 1)$, F is bounded on $U_{\rho-\delta}$ by

$$\|F\|_{\rho-\delta} \le \kappa_0 \, \delta^{-\alpha+1} \|G\|_\rho \; , \qquad \left\| \frac{\partial F}{\partial \theta} \right\|_{\rho-\delta} \le \kappa_1 \, \delta^{-\alpha} \, \|G\|_\rho \; ,$$

where $\alpha = \nu + d + 1$ and $\kappa_0 = \gamma^{-1}\, 8^d \, 2^\nu \nu!$, $\kappa_1 = \gamma^{-1}\, 8^d \, 2^{\nu+1} (\nu+1)!$.

Rüssmann (1975, 1976) has shown that the estimates hold with the optimal exponent $\alpha = \nu + 1$ and with $\kappa_0 = 2^{d+1-\nu} \sqrt{(2\nu)!}$ and $\kappa_1 = 2^{d-\nu} \sqrt{(2\nu+2)!}$. This optimal value of α would yield slightly more favourable estimates in the following, but here we content ourselves with the simpler result given above.

Proof of Lemma 4.1. We have the Fourier series, convergent on the complex extension $\|\mathrm{Im}\,\theta\| < \rho$,

$$G(\theta) - \overline{G} = \sum_{k\neq 0} g_k\, e^{ik\cdot\theta}\ , \qquad F(\theta) = \sum_k f_k\, e^{ik\cdot\theta}$$

with Fourier coefficients $f_0 = \overline{F} = 0$ and

$$f_k = -\frac{g_k}{ik\cdot\omega} \qquad \text{for} \quad k \in \mathbb{Z}^d,\ k \neq 0\ .$$

By Cauchy's estimates, $|g_k| \leq M\, e^{-|k|\rho}$ with $M = \|G - \overline{G}\|_\rho \leq 2\|G\|_\rho$ and $|k| = \sum |k_i|$. It follows with (2.4) that

$$\begin{aligned} \|F\|_{\rho-\delta} &\leq \sum_k |f_k|\, e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma} \sum_k |k|^\nu\, e^{-|k|\delta}\ , \\ \left\|\frac{\partial F}{\partial\theta}\right\|_{\rho-\delta} &\leq \sum_k |f_k|\cdot|k|\, e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma} \sum_k |k|^{\nu+1}\, e^{-|k|\delta}\ . \end{aligned}$$

It remains to bound the right-hand sums. We use the inequality $x^\nu/\nu! \leq e^x$ with $x = |k|\delta/2$ to obtain

$$\sum_k |k|^\nu\, e^{-|k|\delta} \leq 2^\nu \delta^{-\nu} \nu! \sum_k e^{-|k|\delta/2}\ .$$

The last sum is bounded by

$$\sum_k e^{-|k|\delta/2} = \Big(1 + 2\sum_{j=1}^\infty e^{-j\delta/2}\Big)^d = \left(\frac{1+e^{-\delta/2}}{1-e^{-\delta/2}}\right)^d \leq (8\delta^{-1})^d\ .$$

Taken together, the above inequalities yield the stated bound for $\|F\|_{\rho-\delta}$. The bound for the derivative is obtained in the same way, with ν replaced by $\nu+1$. □

The coefficients of the perturbation series (2.5) are bounded as follows.

Lemma 4.2. *Let H_0, H_1 be real-analytic and bounded by M on the complex r-neighbourhood $B_r(b^*)$ of $b^* \in \mathbb{R}^d$ and on $B_r(b^*) \times U_\rho$, respectively. Suppose that $\omega(b^*) = (\partial H_0/\partial a)(b^*)$ satisfies the diophantine condition (2.4). Then, the coefficients of the perturbation series (2.5) are bounded by*

$$\left\|\frac{\partial S_j}{\partial\theta}(b^*,\cdot)\right\|_{\rho/2} \leq C_0 (C_1 j^\alpha)^{j-1}$$

for all $j \geq 0$. Here $C_0 = 2r$, and $C_1 = 128(\kappa_1 M/r\rho^\alpha)^2$ with α and κ_1 of Lemma 4.1.

Proof. We recall from Sect. X.2.2 that S_j is determined by (2.6), where $K_1 = H_1$ and for $j \geq 2$,

$$K_j = \sum_{i=2}^{j} \sum_{k_1+\ldots+k_i=j} \frac{1}{i!} \frac{\partial^i H_0}{\partial a^i} \left(\frac{\partial S_{k_1}}{\partial \theta}, \ldots, \frac{\partial S_{k_i}}{\partial \theta} \right) + \sum_{i=1}^{j-1} \sum_{k_1+\ldots+k_i=j-1} \frac{1}{i!} \frac{\partial^i H_1}{\partial a^i} \left(\frac{\partial S_{k_1}}{\partial \theta}, \ldots, \frac{\partial S_{k_i}}{\partial \theta} \right) .$$

We fix an index, say J, set $\delta = \rho/(2J)$ and abbreviate

$$\|K_k\|_j = \|K_k(b^*, \cdot)\|_{\rho - j\delta}$$

and similarly for $\partial S_k / \partial \theta$. By (2.6) and Lemma 4.1, we have

$$\left\| \frac{\partial S_j}{\partial \theta} \right\|_j \leq \kappa_1 \delta^{-\alpha} \, \|K_j\|_{j-1} .$$

We use the Cauchy estimate

$$\left| \frac{1}{i!} \frac{\partial^i H_0}{\partial a^i} (v_1, \ldots, v_i) \right| \leq \frac{M}{r^i} \, |v_1| \cdot \ldots \cdot |v_i| \, ,$$

where $|\cdot|$ denotes the sum norm on $\mathbb{C}^d$, and bound $\|\cdot\|_{j-1}$ by $\|\cdot\|_k$ for $k \leq j-1$. We thus obtain from the above formula for K_j

$$\|K_j\|_{j-1} \leq \sum_{i=2}^{j} \sum_{k_1+\ldots+k_i=j} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \theta} \right\|_{k_1} \cdot \ldots \cdot \left\| \frac{\partial S_{k_i}}{\partial \theta} \right\|_{k_i} + \sum_{i=1}^{j-1} \sum_{k_1+\ldots+k_i=j-1} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \theta} \right\|_{k_1} \cdot \ldots \cdot \left\| \frac{\partial S_{k_i}}{\partial \theta} \right\|_{k_i} .$$

Combining the two bounds yields

$$\frac{1}{r} \left\| \frac{\partial S_j}{\partial \theta} \right\|_j \leq \beta_j \, ,$$

where, with $\mu = (M/r)(\kappa_1/\delta^\alpha)$, we have $\beta_1 = \mu$ and recursively for $j \geq 2$,

$$\beta_j = \mu \sum_{i=2}^{j} \sum_{k_1+\ldots+k_i=j} \beta_{k_1} \cdot \ldots \cdot \beta_{k_i} + \mu \sum_{i=1}^{j-1} \sum_{k_1+\ldots+k_i=j-1} \beta_{k_1} \cdot \ldots \cdot \beta_{k_i} .$$

Multiplying this equation with ζ^j and summing over j, we see that the generating function $b(\zeta) = \sum_{j=1}^{\infty} \beta_j \zeta^j$ is given implicitly by

$$b(\zeta) - \mu\zeta = \mu \left(\frac{1}{1 - b(\zeta)} - 1 - b(\zeta) \right) + \mu\zeta \left(\frac{1}{1 - b(\zeta)} - 1 \right) ,$$

or explicitly, after solving the quadratic equation, by

$$b(\zeta) = \frac{1}{2}\,\frac{1}{1+\mu} - \sqrt{\tfrac{1}{4}\Big(\frac{1}{1+\mu}\Big)^2 - \frac{\mu}{1+\mu}\,\zeta}\,.$$

Hence, $b(\zeta)$ is analytic on the disc $|\zeta| < 1/(4\mu(1+\mu))$, and is there bounded by $1/(2(1+\mu))$. For $\mu \ge 1$, Cauchy's estimate yields

$$\|\partial S_j/\partial\theta\|_j \le r\beta_j \le 2r\,(8\mu^2)^{j-1}\,.$$

(For the uninteresting case $\mu \le 1$ the bound is $2r\cdot 8^{j-1}$.) For $j = J$ this almost gives the stated result upon inserting the definition of μ, but with an exponent 2α instead of α. This can be reduced to α if in the above proof δ is chosen as $\delta_1 = \rho/4$ in the first step and in the other steps as $\delta_j = \rho/(4J)$. This leads to a more complicated quadratic equation where now $b(\zeta)$ is analytic for $|\zeta| \le (C_1 J^\alpha)^{-1}$. We omit the details of this refinement of the proof. □

For the remainder term in (2.7) we then obtain the following bound.

Lemma 4.3. *In the situation of Lemma 4.2, with* $r \le 1$ *and for* $C_1 N^\alpha \le 1/(2\varepsilon)$*,*

$$\|R_N(b^*,\cdot)\|_{\rho/2} \le 4Mr\Big(\frac{4C_1}{r}\,N^\alpha\Big)^N\,.$$

Proof. The remainder term R_N in (2.7) is a sum of terms

$$\frac{1}{i!}\,\frac{\partial^i H_{k_0}}{\partial a^i}\,(Q_{k_1},\ldots,Q_{k_i}) \qquad \text{for } k_0+k_1+\ldots+k_i = N\,,$$

where

$$Q_k = \frac{\partial S_k}{\partial\theta} + \varepsilon\,\frac{\partial S_{k+1}}{\partial\theta} + \ldots + \varepsilon^{N-k-1}\frac{\partial S_{N-1}}{\partial\theta}\,.$$

As long as $C_1 N^\alpha \le 1/(2\varepsilon)$, we have, by Lemma 4.2,

$$\begin{aligned}\|Q_k(b^*,\cdot)\|_{\rho/2} \;&\le\; \sum_{j=k}^{N-1}\varepsilon^{(j-k)}C_0(C_1 j^\alpha)^j \\ &\le\; C_0\sum_{j=k}^{N-1}2^{-(j-k)}\Big(\frac{j}{N}\Big)^{\alpha j}(C_1N^\alpha)^k \le 2C_0\,(C_1N^\alpha)^k\,.\end{aligned}$$

This implies

$$\Big\|\frac{1}{i!}\,\frac{\partial^i H_{k_0}}{\partial a^i}\,(Q_{k_1},\ldots,Q_{k_i})(b^*,\cdot)\Big\|_{\rho/2} \le \frac{M}{r^i}\,2C_0(C_1N^\alpha)^N$$

for $k_0+k_1+\ldots+k_i = N$. (This bound is also valid when an argument different from b^* appears in the derivatives of H_0 and H_1, as is needed for the remainder terms in the Taylor expansion.) Estimating the number of such expressions by

$$2\sum_{i=1}^{N}\binom{N+i-1}{i} \le 2\sum_{i=0}^{2N-1}\binom{2N-1}{i} = 2^{2N}$$

yields the result. □

X.4.2 Near-Invariant Tori of Perturbed Integrable Systems

The following result extends Lemma 2.1 to exponentially long times for sufficiently small values of the perturbation parameter.

Theorem 4.4. *Let H_0, H_1 be real-analytic on the complex r-neighbourhood $B_r(b^*)$ of $b^* \in \mathbb{R}^d$ and on $B_r(b^*) \times U_\rho$, respectively, with $r \le 1$ and $\rho \le 1$. Suppose that $\omega(b^*) = (\partial H_0/\partial a)(b^*)$ satisfies the diophantine condition (2.4). There are positive constants ε_0, c_0, C such that the following holds for every positive $\beta \le 1$ and for $\varepsilon \le \varepsilon_0$: there exists a real-analytic symplectic change of coordinates $(a,\theta) \mapsto (b,\varphi)$ such that every solution $(b(t),\varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \le c_0\varepsilon^{2\beta}$, satisfies*

$$\|b(t) - b(0)\| \le Ct\exp(-c\,\varepsilon^{-\beta/\alpha}) \quad \textit{for} \quad t \le \exp(\tfrac{1}{2}c\,\varepsilon^{-\beta/\alpha}) .$$

Here, $\alpha = \nu + d + 1$ and $c = (16\,C_1 e/r)^{-1/\alpha}$ with C_1 of Lemma 4.2. Moreover, the transformation is such that, for (a,θ) and (b,φ) related by the above coordinate transform,

$$\|a - b\| \le C\varepsilon \quad \textit{for} \quad \|b - b^*\| \le c_0\,\varepsilon^{2\beta},\ \varphi \in U_{\rho/2} .$$

The thresholds ε_0 and c_0 are such that $\varepsilon_0^{2\beta}$ is inversely proportional to γC_1^2, and c_0 is proportional to γC_1^2.

Remark 4.5. Theorem 4.4 is a *local* result, showing that for b_0 near b^* the tori $\{b = b_0,\ \varphi \in \mathbb{T}^d\}$ are nearly invariant, up to exponentially small deviations, over exponentially long times. Nekhoroshev (1977, 1979) has shown the *global* result, under a “steepness condition” which is in particular satisfied for convex Hamiltonians, that for sufficiently small ε *every* solution of the perturbed Hamiltonian system satisfies, for some positive constants $A, B < 1$ (proportional to the inverse of the square of the dimension),

$$\|a(t) - a(0)\| \le \varepsilon^B \quad \text{for} \quad t \le \exp(\varepsilon^{-A}) .$$

Remark 4.6. The constant C_1 in Lemma 4.2 and constants in similar estimates of Hamiltonian perturbation theory are very large, with the consequence that the results on the long-time behaviour derived from them are meaningful, in a rigorous sense, only for extremely small values of the perturbation parameter ε. Nevertheless, apart from their pure theoretical interest these results are of value as they describe the behaviour to be expected if one presupposes that the constants obtained from the worst-case estimations are unduly pessimistic for a given problem, as is typically the case.

Proof of Theorem 4.4. The proof combines Lemmas 4.2 and 4.3 with the proof of Lemma 2.1. An appropriate choice of the truncation indices N and m then gives the exponential estimates.

As in the proof of Lemma 2.1, we approximate $H_1(b,\theta)$ by a trigonometric polynomial of order m in θ. The error of this approximation is bounded by $\mathcal{O}(e^{-m\rho/2})$ on $B_r(b^*)\times U_{\rho/2}$, which is $\mathcal{O}(e^{-N})$ for the choice $m=2N/\rho$ made below. By the arguments of the proof of Lemma 2.1, the estimates of Lemmas 4.2 and 4.3 (for γ replaced by $\gamma/2$, which increases C_1 to $4C_1$) are then valid in $\mathcal{O}((jm)^{-\alpha})$ and $\mathcal{O}((Nm)^{-\alpha})$ neighbourhoods of b^*: for a sufficiently small constant c^* and with $C_2=16\,C_1/r$,

$$\begin{aligned} \Big\|\frac{\partial S_j}{\partial\theta}(b,\theta)\Big\| &\le C_0(4C_1j^\alpha)^{j-1} \qquad \text{for} \quad \|b-b^*\|\le c^*(jm)^{-\alpha},\ \theta\in U_{\rho/2}\ , \\ |R_N(b,\theta)| &\le 4Mr\,(C_2N^\alpha)^N \qquad \text{for} \quad \|b-b^*\|\le c^*(Nm)^{-\alpha},\ \theta\in U_{\rho/2}\ . \end{aligned}$$

We now consider the symplectic change of variables $(a,\theta)\mapsto(b,\varphi)$ defined by the generating function $S(b,\theta)$. The Hamiltonian equations in the variables (b,φ) are then of the form, for $\|b-b^*\|\le c^*(Nm)^{-\alpha}$,

$$\begin{aligned} \dot b &= -\frac{\partial K}{\partial\varphi}(b,\varphi) = -\varepsilon^N\,\frac{\partial R_N}{\partial\theta}\,\frac{\partial\theta}{\partial\varphi} = \mathcal{O}(\varepsilon^N(C_2N^\alpha)^N) \\ \dot\varphi &= \frac{\partial K}{\partial b}(b,\varphi) = \omega_{\varepsilon,N}(b) + \mathcal{O}((Nm)^\alpha\cdot\varepsilon^N(C_2N^\alpha)^N)\ . \end{aligned} \tag{4.1}$$

Choosing $m=2N/\rho$ and N such that $C_2N^\alpha = 1/(e\varepsilon^\beta)$ gives

$$\begin{aligned} \dot b &= \mathcal{O}(\exp(-c\varepsilon^{-\beta/\alpha})) \\ \dot\varphi &= \omega_{\varepsilon,N}(b) + \mathcal{O}(\varepsilon^{-2\beta}\exp(-c\varepsilon^{-\beta/\alpha})) \end{aligned} \qquad \text{for} \quad \|b-b^*\|\le c_0\,\varepsilon^{2\beta} \tag{4.2}$$

with $c=(C_2e)^{-\alpha}$, which yields the result. □

X.4.3 Near-Invariant Tori of Symplectic Integrators

We return to the situation of Sect. X.3 and apply a symplectic numerical method to the integrable Hamiltonian system (3.1) with (3.2) and (3.3).

Theorem 4.7. *Consider applying a symplectic numerical integrator of order p to the real-analytic completely integrable Hamiltonian system (3.1). Suppose that $\omega(a^*)$ satisfies the diophantine condition (2.4). Then, there exist positive constants c_0, c, C and h_0 such that the following holds for all step sizes $h\le h_0$ and for all $\mu\le\min(p,\alpha)$ with $\alpha=\nu+d+1$: every numerical solution starting with $\|I(p_0,q_0)-a^*\|\le c_0h^{2\mu}$ satisfies*

$$\|I(p_n,q_n)-I(p_0,q_0)\|\le C\,h^p \qquad \textit{for} \quad nh\le\exp(c\,h^{-\mu/\alpha})\ .$$

The constants h_0, c_0, c, C depend on d,γ,ν, on bounds of the real-analytic Hamiltonian H on a complex neighbourhood of the torus $\{(p,q)\,;\,I(p,q)=a^*\}$, and on the numerical method.

Proof. The proof is obtained by following the arguments of the proof of Theorem 3.1. Instead of Lemma 2.1, now Theorem 4.4 is applied to the modified Hamiltonian system (3.5) with $\varepsilon = h^p$. This gives a change of coordinates $(a, \theta) \mapsto (b, \varphi)$ $\mathcal{O}(h^p)$-close to the identity, such that in the new variables, the solution $(\widetilde{b}(t), \widetilde{\varphi}(t))$ of (3.5) satisfies

$$\widetilde{b}(t) = b_0 + \mathcal{O}(\exp(-ch^{-\mu/\alpha})) \qquad \text{for} \quad t \le \exp(c\, h^{-\mu/\alpha}) \,.$$

On the other hand, using the exponentially small bound of Theorem IX.7.6, together with Theorem 4.4 and the arguments of part (b) of the proof of Theorem 3.1, yields for the numerical solution in the new variables

$$b_n = \widetilde{b}(t) + \mathcal{O}(\exp(-ch^{-\mu/\alpha})) \qquad \text{for} \quad t = nh \le \exp(c\, h^{-\mu/\alpha}) \,.$$

Together with $a_n - b_n = \mathcal{O}(h^p)$ this gives the result. □

Remark 4.8. When the symplectic method is applied to a perturbed integrable system as in Theorem 4.4, then the same argument yields for $\|I(p_0, q_0) - a^*\| \le c_0 \eta^{2\beta}$ with $\eta = \max(\varepsilon, h^p)$ and $\beta \le 1$ the bound

$$\|I(p_n, q_n) - I(p_0, q_0)\| \le C\, \eta \qquad \text{for} \quad t \le \exp(c\, \eta^{-\beta/\alpha}) \,.$$

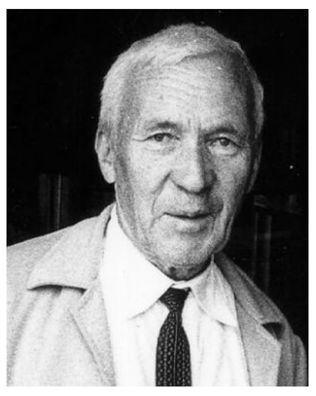
X.5 Kolmogorov's Theorem on Invariant Tori

> (The proof of this theorem was published in Dokl. Akad. Nauk SSSR **98** (1954), 527–530 [MR **16**, 924], but the convergence discussion does not seem convincing to the reviewer.) This very interesting theorem would imply that for an analytic canonical system which is close to an integrable one, all solutions but a set of small measure lie on invariant tori.
>
> (J. Moser 1959)

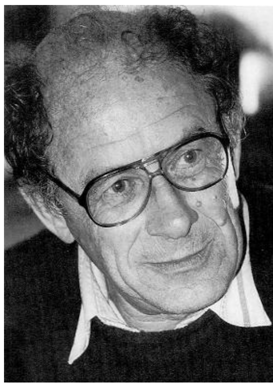
It was a celebrated discovery by Kolmogorov (1954) that invariant tori carrying a conditionally periodic flow with diophantine frequencies persist under small perturbations of the Hamiltonian. Together with the extensions and refinements by Arnold (1963), Moser (1962) and later authors, Kolmogorov's result forms what is now known as KAM theory. Here we give a proof of Kolmogorov's theorem and use it in studying the long-time behaviour of symplectic numerical methods applied to perturbed integrable systems.

X.5.1 Kolmogorov's Theorem

In Sect. X.2.3 we have already given Kolmogorov's transformation which reduces the size of a perturbation to a Hamiltonian of the form (2.8) from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$, at least formally. The iteration of that procedure is convergent and yields the following result.



A.N. Kolmogorov [2]



V.I. Arnold[3]



J.K. Moser [4]

Theorem 5.1 (Kolmogorov 1954). *Consider a real-analytic Hamiltonian $H(a,\theta)$, defined for a in a neighbourhood of $0 \in \mathbb{R}^d$ and $\theta \in \mathbb{T}^d$, for which the linearization at $a^* = 0$ does not depend on the angles:*

$$H(a,\theta) = c + \omega \cdot a + \tfrac{1}{2} a^T M(a,\theta) a \, . \tag{5.1}$$

Suppose that $\omega \in \mathbb{R}^d$ satisfies the diophantine condition (2.4), viz.,

$$|k \cdot \omega| \geq \gamma \, |k|^{-\nu} \quad \text{for} \quad k \in \mathbb{Z}^d, k \neq 0, \tag{5.2}$$

and that the angular average $\overline{M}_0$ of $M(0,\cdot)$ is an invertible $d \times d$ matrix:

$$\|\overline{M}_0 v\| \geq \mu \|v\| \quad \text{for} \quad v \in \mathbb{R}^d, \tag{5.3}$$

with positive constants γ, ν, μ. Let $H_\varepsilon(a,\theta) = H(a,\theta) + \varepsilon G(a,\theta)$ be a real-analytic perturbation of $H(a,\theta)$. Then, there exists $\varepsilon_0 > 0$ such that for every ε with $|\varepsilon| \leq \varepsilon_0$, there is an analytic symplectic transformation $\psi_\varepsilon : (b,\varphi) \mapsto (a,\theta)$, $\mathcal{O}(\varepsilon)$ close to the identity and depending analytically on ε, which puts the perturbed Hamiltonian back to the form

$$H_\varepsilon(a,\theta) = c_\varepsilon + \omega \cdot b + \tfrac{1}{2} b^T M_\varepsilon(b,\varphi) b \quad \text{for} \quad (a,\theta) = \psi_\varepsilon(b,\varphi). \tag{5.4}$$

The perturbed system therefore has the invariant torus $\{b = 0, \ \varphi \in \mathbb{T}^d\}$ carrying a quasi-periodic flow with the same frequencies ω as the unperturbed system.
(The threshold ε_0 depends on d, ν, γ, μ and on bounds of H and G on a complex neighbourhood of $\{0\} \times \mathbb{T}^d$.)

[2] Andrei Nikolaevich Kolmogorov, born: 25 April 1903 in Tambov (Russia), died: 20 October 1987 in Moscow.
[3] Vladimir Igorevich Arnold, born: 12 June 1937 in Odessa (USSR).
[4] Jürgen K. Moser, born: 4 July 1928 in Königsberg, now Kaliningrad, died: 17 December 1999 in Zürich (Switzerland).

Of particular interest is the case when $H(a,\theta) = H_0(a)$ is independent of θ, so that we are considering perturbations of an integrable system. In this case, the theorem shows that all invariant tori with frequencies $\omega(a) = \partial H_0/\partial a(a)$ satisfying (5.2) and with invertible Hessian $\partial^2 H_0/\partial a^2(a)$ persist under small perturbations and are only slightly deformed.

Kolmogorov (1954) stated the theorem and formulated the iteration of Section X.2.3, but did not give the details of the convergence estimates. Arnold (1963) gave a first complete proof of the theorem for perturbed integrable systems, using a construction based on the "ultra-violet cutoff" (cf. Lemma 2.1) which yields a single transformation simultaneously for all frequencies satisfying the diophantine condition (2.4), in contrast to Kolmogorov's iteration which yields a different transformation for every choice of diophantine frequencies. However, Arnold's transformation is no longer analytic in the perturbation parameter ε. Moser (1962) showed that the analyticity of the Hamiltonian can be replaced by differentiability of sufficiently high order. Full proofs of Kolmogorov's theorem along his original construction were published by Thirring (1977) (for a reduced model problem) and by Benettin, Galgani, Giorgilli & Strelcyn (1984).

As in Remark 4.6, a practical difficulty with Theorem 5.1 is that the theoretically obtained threshold ε_0 is very small. The proof below requires $\varepsilon_0 \le \delta_0^{5\alpha}$ with $\alpha = \nu + d + 1$ of Lemma 4.1, where δ_0 is inversely proportional to ν. This pessimistic estimate of the threshold can be somewhat improved by first reducing the perturbation of an integrable Hamiltonian system via a perturbation series expansion as in the proof of Theorem 4.4 and then applying Kolmogorov's theorem to the remainder of the truncated perturbation series.

The proof of Theorem 5.1 uses iteratively the following lemma, which refers to the transformation constructed in Sect. X.2.3. Similar to Sect. X.4 we use the notation

$$\|G\|_\rho = \sup\{|G(a,\theta)|\,;\ \|a\| < \rho,\ \|\mathrm{Im}\,\theta\| < \rho\}$$

for a bounded analytic function G on $W_\rho := B_\rho(0) \times U_\rho$, where again $B_\rho(0)$ is the complex ball of radius ρ around 0 and U_ρ is the complex extension of $\mathbb{T}^d$ of width ρ. The same notation is used for vector- and matrix-valued functions, in which case the underlying norm on $\mathbb{C}^d$ or $\mathbb{C}^{d\times d}$ is the maximum norm or its induced matrix norm, respectively.

Lemma 5.2. *In the situation of Sect. X.2.3 and under the conditions of Theorem 5.1, suppose that H and G are real-analytic and bounded on W_ρ. Then, there exists $\delta_0 > 0$ such that the following bounds hold for Kolmogorov's transformation whenever $0 < \delta \le \delta_0$:*

$$\textit{if}\quad \|\varepsilon G\|_\rho \le \delta^{5\alpha}, \quad \textit{then}\quad \|\varepsilon^2 \widehat{G}\|_{\rho-\delta} \le (\tfrac{1}{2}\delta)^{5\alpha}$$

$$\textit{and}\quad \|\varepsilon \nabla\chi\|_{\rho-\delta} \le \delta^{3\alpha}\,, \quad \|\widehat{M} - M\|_{\rho-\delta} \le \delta^{2\alpha},$$

where $\alpha = \nu + d + 1$. The threshold δ_0 depends only on d, ν, γ, μ and on $\|H\|_\rho$.

Proof. We estimate the terms arising in the construction of Kolmogorov's transformation of Sect. X.2.3. For brevity we denote $\|\cdot\|_j = \|\cdot\|_{\rho - j\delta/4}$ for $j = 0,1,2,3,4$.

(a) The transformation $(b,\varphi) \mapsto (a,\theta)$ is constructed such that $(a,\theta) = y(\varepsilon)$, where $y(t)$ is the solution of $\dot{y} = J^{-1}\nabla\chi(y)$ with $y(0) = (b,\varphi)$. Suppose for the moment that

$$\|\varepsilon\nabla\chi\|_3 \le \tfrac{1}{4}\delta \,. \tag{5.5}$$

Let $(b,\varphi) \in W_{\rho-\delta}$. Then, $\|y(t) - y(0)\| \le \frac{1}{4}\delta$ for $0 \le t \le \varepsilon$, and in particular $\|(a,\theta) - (b,\varphi)\| \le \frac{1}{4}\delta$. We define

$$\begin{aligned} \varepsilon^2 R(b,\varphi) \quad &:= \quad \Big(a - b + \varepsilon\frac{\partial\chi}{\partial\varphi}(b,\varphi),\ \theta - \varphi - \varepsilon\frac{\partial\chi}{\partial b}(b,\varphi)\Big) \\ &= \quad y(\varepsilon) - y(0) - \varepsilon J^{-1}\nabla\chi(y(0)) \end{aligned}$$

and note

$$\|R(b,\varphi)\| \le \tfrac{1}{2}\max_{0\le t\le\varepsilon}\|\ddot{y}(t)\| \le \tfrac{1}{2}\|J^{-1}\nabla^2\chi\ J^{-1}\nabla\chi\|_3$$

so that

$$\|R\|_4 \le \tfrac{1}{2}\|\nabla^2\chi\|_3\,\|\nabla\chi\|_3 \,. \tag{5.6}$$

(b) Tracing the construction of Sect. X.2.3, we find by Taylor expansion of $H(a,\theta)$ that the new matrix is

$$\widehat{M}(b,\varphi) = M(b,\varphi) + \varepsilon L(b,\varphi)$$

with

$$L(b,\varphi) = \sum_{i=1}^{d}\left(\frac{\partial M}{\partial a_i}\,\frac{\partial\chi}{\partial\varphi_i} - \frac{\partial M}{\partial\theta_i}\,\frac{\partial\chi}{\partial b_i}\right)(b,\varphi) + P(b,\varphi) + Q(b,\varphi)$$

where $P(b,\varphi)$ is symmetric with

$$b^T P(b,\varphi) b = b^T\big(M(b,\varphi) - M(0,\varphi)\big)\frac{\partial\chi}{\partial\varphi}$$

and where $Q(b,\varphi)$ is given by (2.11). It follows that

$$\|\widehat{M} - M\|_4 \le 2\varepsilon(\|\nabla M\|_4\,\|\nabla\chi\|_4 + \|\nabla^2 G\|_4) \,. \tag{5.7}$$

From the construction of $\widehat{G}$ we also find by simple estimates of Taylor remainders

$$\|\widehat{G}\|_4 \le \|\nabla H\|_3\,\|R\|_4 + \|\nabla G\|_3\,\|\nabla\chi\|_4 + \|\nabla^2 H\|_3\,\|\nabla\chi\|_4^2 \,. \tag{5.8}$$

(c) Using Lemma 4.1 in the equations (2.12)–(2.16) defining χ of (2.10), we obtain first

$$\|\chi_0\|_1 \le \kappa_0\delta^{-\alpha+1}\|G_0\|_0 \,, \qquad \Big\|\frac{\partial\chi_0}{\partial\varphi}\Big\|_1 \le \kappa_1\delta^{-\alpha}\|G_0\|_0$$

and by a second application of that lemma, for $i = 1, \dots, d$,

$$\|\chi_i\|_2 \le \kappa_0 \delta^{-\alpha+1} \left(\|u\|_1 + \|v\|_1 + \|G_i\|_1\right)$$

where, by construction of u and v,

$$\|v\|_1 \le \|M\|_1 \left\|\frac{\partial \chi_0}{\partial \varphi}\right\|_1, \qquad \|u\|_1 \le \|M\|_1 \, \mu^{-1}\Big(\|v\|_1 + \sum_{j=1}^{d} \|G_j\|_1\Big) .$$

It then follows by Cauchy's estimates that

$$\|\nabla \chi\|_3 \le C\delta^{-2\alpha} \, \|G\|_0 \,, \qquad \|\nabla^2 \chi\|_3 \le C\delta^{-2\alpha-1} \, \|G\|_0 \,. \tag{5.9}$$

(d) Combining the estimates (5.6)–(5.9) and using once more Cauchy's estimates to bound derivatives of H and G yields

$$\begin{aligned} \|\varepsilon^2 \widehat{G}\|_{\rho-\delta} &\le C\delta^{-4\alpha-1} \|\varepsilon G\|_\rho^2 \\ \|\varepsilon \nabla \chi\|_{\rho-\delta} &\le C\delta^{-2\alpha} \|\varepsilon G\|_\rho \\ \|\widehat{M} - M\|_{\rho-\delta} &\le C\delta^{-2\alpha-3} \|\varepsilon G\|_\rho \,. \end{aligned}$$

All this holds under the condition (5.5). By (5.9), this condition is satisfied if $\|\varepsilon G\|_\rho \le \delta^{5\alpha}$ and $\delta \le \delta_0$ with a sufficiently small δ_0. (Tracing the above constants shows that δ_0 needs to be inversely proportional to $\kappa_1^{1/\alpha}$, or inversely proportional to ν.) This yields the stated bounds. □

Proof of Theorem 5.1. Kolmogorov's iteration yields sequences

$$\begin{aligned} &G^{(0)} = G, G^{(1)}, G^{(2)}, \dots \\ &M^{(0)} = M, M^{(1)}, M^{(2)}, \dots \\ &\chi^{(0)}, \chi^{(1)}, \chi^{(2)}, \dots \,. \end{aligned}$$

By Lemma 5.2 they satisfy, provided that $\|\varepsilon G\|_\rho = \delta^{5\alpha}$ with $\delta \le \delta_0$,

$$\|\varepsilon^{2^j} G^{(j)}\|_{\rho^{(j)}} \le (2^{-j}\delta)^{5\alpha} \tag{5.10}$$

$$\|M^{(j+1)} - M^{(j)}\|_{\rho^{(j)}} \le (2^{-j}\delta)^{2\alpha} \tag{5.11}$$

$$\|\varepsilon^{2^j} \nabla \chi^{(j)}\|_{\rho^{(j)}} \le (2^{-j}\delta)^{3\alpha} \tag{5.12}$$

where $\rho^{(j)} = \rho - (1 + \frac{1}{2} + \ldots + 2^{-j})\delta > \frac{1}{2}\rho$ for all j. Note that (5.11) implies that the inverse of $M^{(j)}$ is bounded by $2\mu^{-1}$ for all j, so that the iterative use of Lemma 5.2 is justified. The time-ε^{2^j} flow of $\chi^{(j)}$ is a symplectic transformation $\sigma_\varepsilon^{(j)}$, which by (5.12) satisfies

$$\|\sigma_\varepsilon^{(j)} - \mathrm{id}\|_{\rho/2} \le (2^{-j}\delta)^{3\alpha} \,. \tag{5.13}$$

The composed transformation

$$\psi_\varepsilon^{(j)} := \sigma_\varepsilon^{(0)} \circ \sigma_\varepsilon^{(1)} \circ \ldots \circ \sigma_\varepsilon^{(j)}$$

is constructed such that

$$H(\psi_\varepsilon^{(j-1)}(b,\varphi)) = c^{(j)} + \omega \cdot b + b^T M^{(j)}(b,\varphi) b + \varepsilon^{2^j} G^{(j)}(b,\varphi) \; . \tag{5.14}$$

By (5.13), the sequence $\psi_\varepsilon^{(j)}(b,\varphi)$ converges uniformly on $W_{\rho/2} \times (-\varepsilon_0, \varepsilon_0)$ to a limit $\psi_\varepsilon(b,\varphi)$. By Weierstrass' theorem, $\psi_\varepsilon(b,\varphi)$ is analytic in (b,φ,ε) (and in any further parameters on which M and G might possibly depend analytically). Since ψ_ε depends analytically on ε and $\psi_0 = \text{id}$, it follows that ψ_ε is $\mathcal{O}(\varepsilon)$-close to the identity on $W_{\rho/2}$. By (5.10) and (5.14), the transformed Hamiltonian $H \circ \psi_\varepsilon$ is of the desired form (5.4). □

X.5.2 KAM Tori under Symplectic Discretization

Consider a Hamiltonian system

$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial q}(p,q) \; , \quad \dot{q} = \frac{\partial \mathcal{H}}{\partial p}(p,q) \; , \tag{5.15}$$

for which, in suitable coordinates (a,θ), the Hamiltonian $\mathcal{H}(p,q) = H(a,\theta) + \varepsilon G(a,\theta)$ satisfies the conditions of Theorem 5.1. Kolmogorov's theorem yields a transformation to variables (b,φ) in terms of which

$$\mathcal{H}(p,q) = \omega \cdot b + \tfrac{1}{2} b^T M_\varepsilon(b,\varphi) b \; ,$$

so that the torus $\mathcal{T}_\omega = \{b = 0, \; \varphi \in \mathbb{T}^d\}$ is invariant and the flow on it is quasi-periodic with frequencies ω.

For a symplectic integrator of order p applied to (5.15), backward analysis gives a modified Hamiltonian $\widetilde{\mathcal{H}}(p,q)$ which is an $\mathcal{O}(h^p)$ perturbation of $\mathcal{H}(p,q)$:

$$\widetilde{\mathcal{H}}(p,q) = \omega \cdot b + \tfrac{1}{2} b^T M_\varepsilon(b,\varphi) b + h^p \widetilde{G}(b,\varphi) \; . \tag{5.16}$$

Kolmogorov's theorem can be applied once more, yielding an invariant torus $\widetilde{\mathcal{T}}_\omega$ of the modified Hamiltonian $\widetilde{\mathcal{H}}(p,q)$ which again carries a quasi-periodic flow with frequencies ω. Combined with the exponentially small estimates of backward analysis for the difference between numerical solutions and the flow of the modified Hamiltonian system, this gives the following result of Hairer & Lubich (1997).

Theorem 5.3. *In the above situation, for a symplectic integrator of order p used with sufficiently small step size h, there is a modified Hamiltonian $\widetilde{\mathcal{H}}$ with an invariant torus $\widetilde{\mathcal{T}}_\omega$ carrying a quasi-periodic flow with frequencies ω, $\mathcal{O}(h^p)$ close to the invariant torus $\mathcal{T}_\omega$ of the original Hamiltonian $\mathcal{H}$, such that the difference between any numerical solution (p_n, q_n) starting on the torus $\widetilde{\mathcal{T}}_\omega$ and the solution*

$(\widetilde{p}(t),\widetilde{q}(t))$ of the modified Hamiltonian system with the same starting values remains exponentially small in $1/h$ over exponentially long times:

$$\|(p_n,q_n)-(\widetilde{p}(t),\widetilde{q}(t))\|\le Ce^{-\kappa/h}\qquad \textit{for}\quad t=nh\le e^{\kappa/h}\;.$$

The constants C and κ are independent of n,h,ε (for h,ε sufficiently small) and of the initial value $(p_0,q_0)\in\widetilde{\mathcal{T}}_\omega$.

Proof. (a) For sufficiently small h, Kolmogorov's theorem applied to (5.16) yields a change of coordinates $(b,\varphi)\mapsto(c,\psi)$, $O(h^p)$ close to the identity, which transforms the modified Hamiltonian to the form

$$\widetilde{\mathcal{H}}(p,q)=\omega\cdot c+\tfrac{1}{2}c^TM_{\varepsilon,h}(c,\psi)c\;,$$

with the invariant torus $\widetilde{\mathcal{T}}_\omega=\{c=0,\;\psi\in\mathbb{T}^d\}$. The corresponding differential equations read in these coordinates

$$\dot c=u(c,\psi)\;,\qquad \dot\psi=\omega+v(c,\psi) \tag{5.17}$$

where $u(c,\psi)=\mathcal{O}(\|c\|^2)$ and $v(c,\psi)=\mathcal{O}(\|c\|)$, and similarly for the derivatives $\partial u/\partial c=\mathcal{O}(\|c\|)$, $\partial u/\partial\psi=\mathcal{O}(\|c\|^2)$, and $\partial v/\partial c=\mathcal{O}(1)$, $\partial v/\partial\psi=\mathcal{O}(\|c\|)$. The constants in these $\mathcal{O}$-terms are independent of h and ε. Let $(c(t),\psi(t))$ and $(\widehat{c}(t),\widehat{\psi}(t))$ be two solutions of (5.17) such that $\|c(t)\|\le\beta$, $\|\widehat{c}(t)\|\le\beta$ (β sufficiently small) for all t under consideration. Then, an argument based on Gronwall's lemma shows that their difference is bounded over a time interval $0\le t\le 1/\beta$ by

$$\begin{aligned}\|c(t)-\widehat{c}(t)\|&\le C\left(\|c(0)-\widehat{c}(0)\|+\beta\,\|\psi(0)-\widehat{\psi}(0)\|\right)\\ \|\psi(t)-\widehat{\psi}(t)\|&\le C\left(t\,\|c(0)-\widehat{c}(0)\|+\|\psi(0)-\widehat{\psi}(0)\|\right)\,,\end{aligned}\tag{5.18}$$

for some constant C that does not depend on β, h or ε.

(b) In the following we denote $y=(p,q)$ for brevity, and more specifically, y_n denotes the numerical solution starting from any y_0 on the torus $\widetilde{\mathcal{T}}_\omega$, i.e., the c-coordinate of y_0 vanishes: $c_0=0$. We denote by $\widetilde{y}(t,s,z)$ the solution of the modified Hamiltonian system with initial value $\widetilde{y}(s,s,z)=z$, and more briefly $\widetilde{y}(t)=\widetilde{y}(t,0,y_0)$ the solution starting from y_0. By Theorem IX.7.6, the local error of backward error analysis at $t_j=jh$ is bounded by

$$\|y_j-\widetilde{y}(t_j,t_{j-1},y_{j-1})\|\le\delta:=\textit{Const.}\,h\,e^{-3\kappa/h}$$

for some constant κ, as long as y_j remains in a compact subset of the domain of analyticity of $\mathcal{H}$. We further denote the c-coordinates of y_n, $\widetilde{y}(t)$ and $\widetilde{y}(t,t_j,y_j)$ by c_n, $\widetilde{c}(t)$ and $\widetilde{c}(t,t_j,y_j)$, respectively. To apply the error propagation estimate (5.18), we assume that

$$\|\widetilde{c}(t,t_j,y_j)\|\le\beta\qquad\text{for}\quad t_j\le t\le 1/\beta \tag{5.19}$$

and for all j satisfying $t_j=jh\le 1/\beta$. This assumption will be justified by induction later, and the value of β will be specified in (5.21) below. By (5.18) we thus obtain the bound

$$\|\widetilde{y}(t,t_j,y_j) - \widetilde{y}(t,t_{j-1},y_{j-1})\| \le C\,(1+(t-t_j))\,\delta \qquad \text{for} \quad t_j \le t \le 1/\beta \;.$$

Summing up from $j = 1$ to n gives for $t_n \le t \le 1/\beta$ (and $t > 2$)

$$\begin{aligned} \|\widetilde{y}(t,t_n,y_n) - \widetilde{y}(t)\| \;&\le\; \sum_{j=1}^{n} C\,\big(1+(t-t_j)\big)\,\delta \le Ch^{-1}\delta\,(t_n + tt_n - t_n^2/2) \\ &<\; Ch^{-1}\delta\, t^2 \le Ch^{-1}\delta/\beta^2 \;. \end{aligned} \tag{5.20}$$

We now set

$$\beta = (2Ch^{-1}\delta)^{1/3}\;, \tag{5.21}$$

so that $Ch^{-1}\delta/\beta^2 = \beta/2$, and we obtain the desired estimate from (5.20) by putting $t = t_n$.

(c) We still have to justify the assumption (5.19). This will be done by induction. For $j = 0$ nothing needs to be shown, because $\widetilde{c}(t,0,y_0) = \widetilde{c}(t) \equiv 0$ as a consequence of the fact that $\widetilde{y}(t)$ stays on the invariant torus $\widetilde{\mathcal{T}}_\omega = \{c = 0,\ \psi \in \mathbb{T}^d\}$. Suppose now that (5.19) holds for $j \le n$. It then follows from (5.20) that

$$\|\widetilde{c}(t,t_n,y_n)\| < Ch^{-1}\delta/\beta^2 = \beta/2 \qquad \text{for} \quad t_n \le t \le 1/\beta$$

(again because of $\widetilde{c}(t) \equiv 0$). Consequently we also have

$$\|c_{n+1}\| \le \|c_{n+1} - \widetilde{c}(t_{n+1},t_n,y_n)\| + \|\widetilde{c}(t_{n+1},t_n,y_n)\| < \delta + \beta/2 \le \beta,$$

provided that h is sufficiently small so that $\delta \le \beta/2$. By continuity, $\widetilde{c}(t,t_{n+1},y_{n+1})$ is bounded by β on a non-empty interval $[t_{n+1},T_{n+1}]$. The computation of part (b) shows that $\|\widetilde{c}(t,t_{n+1},y_{n+1})\| \le \beta/2$ on this interval. Hence, T_{n+1} can be increased until $T_{n+1} \ge 1/\beta$. This proves the estimate (5.19) for $j = n+1$. □

X.6 Invariant Tori of Symplectic Maps

In the preceding section, backward error analysis combined with Kolmogorov's theorem has shown that a symplectic integrator applied to a Hamiltonian system with KAM tori possesses tori that are near-invariant, up to exponentially small terms, over exponentially long times in the inverse of the step size. To obtain truly invariant tori, we need a discrete KAM theorem for perturbations of integrable near-identity maps depending on a small parameter, the step size. Such a result was recently obtained by Shang (1999, 2000), who gave a discrete Arnold-type construction. Here, we use instead a discrete-time version of Kolmogorov's iteration. This establishes the existence of invariant tori of symplectic integrators applied to integrable Hamiltonian systems or to near-integrable systems with KAM tori, for a Cantor set of non-resonant step sizes.

X.6.1 A KAM Theorem for Symplectic Near-Identity Maps

We consider a discrete-time analogue of the situation in Sections X.2.3 and X.5.1 and construct the corresponding version of Kolmogorov's iteration. Consider the symplectic map $\sigma_h : (a, \theta) \mapsto (\widehat{a}, \widehat{\theta})$ for a near $0 \in \mathbb{R}^d$, $\theta \in \mathbb{T}^d$ defined by

$$\widehat{a} = a - h\,\frac{\partial S}{\partial \widehat{\theta}}(a, \widehat{\theta})\ , \qquad \widehat{\theta} = \theta + h\,\frac{\partial S}{\partial a}(a, \widehat{\theta}) \tag{6.1}$$

where h is a small parameter (the step size), and $S : B_r(0) \times \mathbb{T}^d \to \mathbb{R}$ is a real-analytic generating function. If $S(a, \widehat{\theta})$ has the form (cf. (2.8))

$$S(a, \widehat{\theta}) = c + \omega \cdot a + \tfrac{1}{2} a^T M(a, \widehat{\theta}) a\ , \tag{6.2}$$

then the associated symplectic map is of the form

$$\widehat{a} = a + \mathcal{O}(h\|a\|^2)\ , \qquad \widehat{\theta} = \theta + h\omega + \mathcal{O}(h\|a\|)\ .$$

Hence, the torus $\{a = 0,\ \theta \in \mathbb{T}^d\}$ is invariant, and on it the map σ_h reduces to rotation by $h\omega$.

Consider now an analytic perturbation of such a generating function: $S(a, \widehat{\theta}) + \varepsilon R(a, \widehat{\theta})$ with a small ε. We construct a near-identity symplectic change of coordinates, via an iterative procedure similar to Kolmogorov's iteration of Sect. X.2.3, such that the generating function of the perturbed symplectic map in the new variables is again of the form (6.2) with the same ω, and hence the perturbed map has an invariant torus on which it is conjugate to rotation by $h\omega$. This holds if $h\omega$ satisfies the following diophantine condition (cf. (2.4)):

$$\left| \frac{1 - e^{-ik\cdot h\omega}}{h} \right| \geq \gamma^*\, |k|^{-\nu^*} \qquad \text{for} \quad k \in \mathbb{Z}^d, k \neq 0, \tag{6.3}$$

for some positive constants γ^*, ν^*; and if the angular average $\overline{M}_0$ of $M(0, \cdot)$ is invertible:

$$\|\overline{M}_0 v\| \geq \mu^* \|v\| \qquad \text{for} \quad v \in \mathbb{R}^d \tag{6.4}$$

for a positive constant μ^*. As in Sect. X.2.3, we construct a symplectic transformation $(a, \theta) \mapsto (b, \varphi)$ as the time-ε flow of an auxiliary Hamiltonian of the form (2.10), viz.,

$$\chi(b, \varphi) = \xi \cdot \varphi + \chi_0(\varphi) + \sum_{i=1}^{d} b_i \chi_i(\varphi)$$

where $\xi \in \mathbb{R}^d$ is a constant vector, and $\chi_0, \chi_1, \ldots, \chi_d$ are 2π-periodic functions. We then consider the map conjugate to the perturbed map $(a, \theta) \mapsto (\widehat{a}, \widehat{\theta})$ generated by $S(a, \widehat{\theta}) + \varepsilon R(a, \widehat{\theta})$:

$$\begin{array}{ccc} (a, \theta) & \longrightarrow & (\widehat{a}, \widehat{\theta}) \\ \uparrow & & \downarrow \\ (b, \varphi) & & (\widehat{b}, \widehat{\varphi}) \end{array}$$

We construct χ in such a way that the above composed symplectic map is generated by $\widetilde{S}(b\widehat{\varphi}) + \varepsilon^2 \widetilde{R}(b, \widehat{\varphi})$ with $\widetilde{S}$ of the form (6.2) and both $\widetilde{S}$ and $\widetilde{R}$ real-analytic and bounded independently of ε and of h with (6.3). The map $(b, \varphi) \mapsto (\widehat{b}, \widehat{\varphi})$ is then of the form

$$\widehat{b} = b + \mathcal{O}(h\|b\|^2) + \mathcal{O}(h\varepsilon^2) \,, \qquad \widehat{\varphi} = \varphi + h\omega + \mathcal{O}(h\|b\|) + \mathcal{O}(h\varepsilon^2) \,.$$

As an elementary calculation shows, this holds if χ satisfies for all $(b, \widehat{\varphi})$ with b near 0, $\widehat{\varphi} \in \mathbb{T}^d$

$$\frac{\chi(b, \widehat{\varphi}) - \chi(b, \widehat{\varphi} - h\omega)}{h} + b^T M(b, \widehat{\varphi}) \frac{\partial \chi}{\partial \varphi}(b, \widehat{\varphi} - h\omega) + R(b, \widehat{\varphi}) = C_h + \mathcal{O}(\|b\|^2)$$

where C_h does not depend on $(b, \widehat{\varphi})$ and ε. Writing down the Taylor expansion

$$R(b, \widehat{\varphi}) = R_0(\widehat{\varphi}) + \sum_{i=1}^{d} b_i R_i(\widehat{\varphi}) + \mathcal{O}(\|b\|^2)$$

and inserting the above ansatz for χ, this condition becomes fulfilled if, with $u(\widehat{\varphi}) = M(0, \widehat{\varphi})\xi$ and $v(\widehat{\varphi}) = M(0, \widehat{\varphi})(\partial \chi_0 / \partial \varphi)(\widehat{\varphi} - h\omega)$,

$$\frac{\chi_0(\widehat{\varphi}) - \chi_0(\widehat{\varphi} - h\omega)}{h} + R_0(\widehat{\varphi}) = \overline{R}_0 \tag{6.5}$$

$$\frac{\chi_i(\widehat{\varphi}) - \chi_i(\widehat{\varphi} - h\omega)}{h} + u_i(\widehat{\varphi}) + v_i(\widehat{\varphi}) + R_i(\widehat{\varphi}) = \overline{u}_i + \overline{v}_i + \overline{R}_i \tag{6.6}$$

$$\overline{u}_i + \overline{v}_i + \overline{R}_i = 0 \qquad (i = 1, \ldots, d) \tag{6.7}$$

where the bars again denote angular averages. We note

$$\frac{\chi_0(\widehat{\varphi}) - \chi_0(\widehat{\varphi} - h\omega)}{h} = \sum_k \frac{1 - e^{-ik\cdot h\omega}}{h} \, \chi_{0,k} \, e^{ik\cdot\widehat{\varphi}} \,,$$

where $\chi_{0,k}$ are the Fourier coefficients of χ_0. Under the diophantine condition (6.3), Equation (6.5) is thus solved like (2.14) under condition (2.4). Equations (6.6) are of the same type. The above system is then solved in the same way as (2.12)–(2.16), yielding that the perturbed map in the new coordinates, $(b, \varphi) \mapsto (\widehat{b}, \widehat{\varphi})$, is generated by

$$S^{(1)}(b, \widehat{\varphi}) = c^{(1)} + \omega \cdot b + \tfrac{1}{2} b^T M^{(1)}(b, \widehat{\varphi}) b + \varepsilon^2 R^{(1)}(b, \widehat{\varphi})$$

with unchanged frequencies ω and with $M^{(1)}(b, \widehat{\varphi}) = M(b, \widehat{\varphi}) + \mathcal{O}(\varepsilon)$. The perturbation to the form (6.2) is thus reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. By the same arguments as in the proof of Theorem 5.1 it is shown that the iteration of this procedure converges. This proves the following discrete-time version of Kolmogorov's theorem.

Theorem 6.1. *Consider a real-analytic function $S(a, \widehat{\theta})$ of the form (6.2) with (6.4), defined on a neighbourhood of $\{0\} \times \mathbb{T}^d$. Let $|h| < h_0$ (h_0 so small that (6.1) is a well-defined map) and suppose that $h\omega$ satisfies (6.3).*

Let $S_\varepsilon(a,\widehat{\theta}) = S(a,\theta) + \varepsilon R(a,\widehat{\theta})$ *be an analytic perturbation of* $S(a,\theta)$*, generating a symplectic map* $\sigma_{h,\varepsilon} : (a,\theta) \mapsto (\widehat{a},\widehat{\theta})$ *via (6.1) with* S_ε *in place of* S*.*

Then, there exists $\varepsilon_0 > 0$ *such that for every* ε *with* $|\varepsilon| < \varepsilon_0$*, there is an analytic symplectic transformation* $\psi_{h,\varepsilon} : (b,\varphi) \mapsto (a,\theta)$*,* $\mathcal{O}(\varepsilon)$ *close to the identity uniformly in* h *satisfying (6.3) and analytic in* ε*, such that* $\psi_{h,\varepsilon}^{-1} \circ \sigma_{h,\varepsilon} \circ \psi_{h,\varepsilon} : (b,\varphi) \mapsto (\widehat{b},\widehat{\varphi})$ *is generated, via (6.1), by a function* $S^*_{h,\varepsilon}(b,\widehat{\varphi})$ *which is again of the form (6.2), i.e.,*

$$S^*_{h,\varepsilon}(b,\widehat{\varphi}) = c_{h,\varepsilon} + \omega \cdot b + \tfrac{1}{2} b^T M_{h,\varepsilon}(b,\widehat{\varphi}) b \,.$$

The perturbed map $\sigma_{h,\varepsilon}$ *therefore has an invariant torus on which it is conjugate to rotation by* $h\omega$*.*
(The threshold ε_0 *depends only on* $d, \nu^*, \gamma^*, \mu^*$ *and on bounds of* S *and* R *on a complex neighbourhood of* $\{0\} \times \mathbb{T}^d$*.)* □

X.6.2 Invariant Tori of Symplectic Integrators

As a direct consequence of Theorem 6.1 we obtain the following result on invariant tori of symplectic integrators applied to KAM systems.

Theorem 6.2. *Apply a symplectic integrator of order* p *to a perturbed integrable system with a KAM torus* $\mathcal{T}_\omega$ *which carries a quasi-periodic flow with diophantine frequencies* ω*. Then, if the step size* h *is sufficiently small and satisfies the strong non-resonance condition (6.3), the numerical method has an invariant torus* $\mathcal{T}_{\omega,h}$ $\mathcal{O}(h^p)$*-close to* $\mathcal{T}_\omega$*, on which it is conjugate to rotation by* $h\omega$*.*

Proof. Theorem 6.1 applies directly, with $\varepsilon = h^p$, to the above situation. Here, the generating function $S(a,\widehat{\theta})$ of the time-h flow φ_h of the Hamiltonian system with the KAM torus $\mathcal{T}_\omega$ is of the form (6.2) in the variables (a,θ) obtained by Kolmogorov's theorem. The matrix $M(a,\widehat{\theta})$ in (6.2) then differs from the corresponding matrix of (2.8) by $\mathcal{O}(h)$, so that (5.3) implies (6.4). Finally, the generating function of the numerical one-step map Φ_h is an $\mathcal{O}(h^p)$-perturbation $S(a,\widehat{\theta}) + h^p R(a,\widehat{\theta})$. □

X.6.3 Strongly Non-Resonant Step Sizes

Theorem 6.2 leaves us with an interesting question: if $\omega \in \mathbb{R}^d$ is a vector of frequencies that satisfies the diophantine condition (2.4), then which step sizes h satisfy the non-resonance condition (6.3)? Here we give a lemma in the spirit of results by Shang (2000). It shows that the probability of picking an $h \in (0,h_0)$ satisfying (6.3) tends to 1 as $h_0 \to 0$.

Lemma 6.3. *Suppose* $\omega \in \mathbb{R}^d$ *satisfies (2.4), and let* $h_0 > 0$*. For any choice of positive* γ^* *and* ν^**, the set*

$$Z(h_0) = \{h \in (0,h_0)\ ;\ \ h \textit{ does } \text{not } \textit{satisfy (6.3)}\}$$

is open and dense in $(0, h_0)$. *If* $\gamma^* \leq \gamma$ *and* $\nu^* > \nu + d + r$ *with* $r > 1$, *then the Lebesgue measure of* $Z(h_0)$ *is bounded by*

$$\text{measure}\,\big(Z(h_0)\big) \leq C\,\frac{\gamma^*}{\gamma}\,h_0^{r+1}$$

where C *depends only on* d, ν, ν^* *and* $\|\omega\|$.

Proof. It is clear from the definition that $Z(h_0)$ is open and dense in $(0, h_0)$. It remains to prove the estimate of the Lebesgue measure. For every $k \in \mathbb{Z}^d$ and $|h| \leq h_0$, there exists an integer $l = l(k, h)$ such that

$$|1 - e^{-ik\cdot h\omega}| \geq \frac{2}{\pi}\,|k \cdot h\omega - 2\pi l| = \frac{2}{\pi}\,|k \cdot \omega| \cdot \Big|h - \frac{2\pi l}{|k \cdot \omega|}\Big|\,.$$

For this l we must have, by the triangle inequality,

$$2\pi |l| \leq \pi + |k|\,h_0\,\|\omega\|\,,$$

so that in case $l \neq 0$

$$\frac{1}{|k|} \leq \frac{h_0\,\|\omega\|}{2\pi(|l| - \frac{1}{2})}\,.$$

On the other hand, $l = 0$ yields

$$\left|\frac{1 - e^{-ik\cdot h\omega}}{h}\right| \geq \frac{2}{\pi}\,|k \cdot \omega| \geq \frac{2}{\pi}\,\gamma\,|k|^{-\nu}$$

which implies $h \notin Z(h_0)$. Hence, h can be in $Z(h_0)$ only if there exist $k \in \mathbb{Z}^d$, $k \neq 0$ and an integer $l \neq 0$ such that

$$\begin{aligned}
\Big|h - \frac{2\pi l}{|k \cdot \omega|}\Big| &\leq \frac{\pi}{2}\,\frac{|h|}{|k \cdot \omega|}\,\frac{\gamma^*}{|k|^{\nu^*}} \leq \frac{\pi}{2}\,|h|\,\frac{|k|^{\nu}}{\gamma}\,\frac{\gamma^*}{|k|^{\nu^*}} \\
&\leq \frac{\pi}{2}\,\frac{\gamma^*}{\gamma}\,|k|^{\nu+r-\nu^*}\,\left(\frac{\|\omega\|}{2\pi}\,\frac{1}{|l| - \frac{1}{2}}\right)^r h_0^{r+1}\,.
\end{aligned}$$

It follows that

$$\text{measure}\,\big(Z(h_0)\big) \leq 2\sum_{k\neq 0}\sum_{l\neq 0}\frac{\pi}{2}\,\frac{\gamma^*}{\gamma}\,|k|^{\nu+r-\nu^*}\,\left(\frac{\|\omega\|}{2\pi}\,\frac{1}{|l| - \frac{1}{2}}\right)^r h_0^{r+1}\,,$$

which yields the stated result. □

X.7 Exercises

1. Let R be a $d \times 2d$ matrix of rank d. Show that there exists a symplectic $2d \times 2d$ matrix A such that $RA = (P, Q)$ with an invertible $d \times d$ matrix P.
   Hint. Consider first the case $d = 2$ and then reduce the general situation to a sequence of transformations for that case.

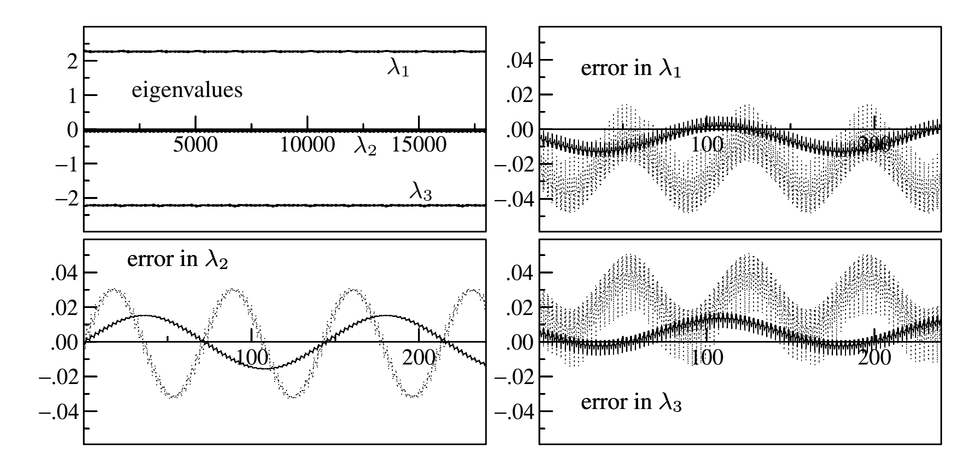


Fig. 7.1. Numerically obtained eigenvalues (left pictures) and errors in the eigenvalues (right pictures) for the step sizes $h = 0.1$ (dotted) and $h = 0.05$ (solid line)

2. The transformation $(x, y) \mapsto (x, y + d(x, y))$ is symplectic if and only if the partial derivatives of d satisfy $d_x = d_x^T\,,\ d_y = 0\,.$
3. In the situation of Lemma 1.1, if $(F_1, \ldots, F_d, \widetilde{G}_1, \ldots, \widetilde{G}_d)^T$ is another such symplectic transformation, then there exists a smooth function W depending only on $x = (x_1, \ldots, x_d)$ such that, for $x_j = F_j(p, q)$,
$$\widetilde{G}_i(p, q) - G_i(p, q) = \frac{\partial W}{\partial x_i}(x)\ .$$
Hint. Use the previous exercise.
4. Show that every discrete subgroup of $\mathbb{R}^d$ is a grid, generated by $k \le d$ linearly independent vectors.
Solution. See e.g. Arnold (1989), Sect. 10D.
5. Show the following bound of the Lebesgue measure of non-diophantine frequencies (Arnold 1963): for any bounded domain $\Omega \subset \mathbb{R}^d$,
$$\text{measure}\big\{\omega \in \Omega\ ;\ \omega \text{ does not satisfy (2.4) with } \nu \ge d\big\} \le C(d, \Omega)\gamma\ .$$
Hint. For a fixed k, decompose $\omega = \omega_0 + \alpha k/|k|$ with $\omega_0 \cdot k = 0$.
6. Show that the eigenvalues λ_j of the matrix L of the Toda system are first integrals in involution.
Hint. For $P_\lambda = \det(\lambda I - L)$, show that $\{P_\lambda, P_\mu\} = 0$ for all λ, μ.
7. We repeat the experiment of Fig. 1.3 with the Störmer–Verlet scheme, where we keep the initial values for the q-variables, but change the initial values for the p-variables to $p_1 = p_2 = p_3 = 0$. The numerical results, given in Fig. 7.1, are qualitatively different from those in Fig. 1.3. The errors behave more like $hc(th)$ rather than $h^2c(t)$. We do not understand this behaviour; do you?
8. Show that for a non-symplectic numerical method, there is at worst quadratic error growth in time when it is applied to an integrable Hamiltonian system.

9. Consider a numerical integrator of order p (i.e., $\Phi_h(y) = \varphi_h(y) + \mathcal{O}(h^{p+1})$), and assume that
$$\Phi_h'(y)^T J \, \Phi_h'(y) \; = \; J + \mathcal{O}(h^{q+1})$$
with $q > p$, when the method is applied to a Hamiltonian system. Prove that under the assumptions of Theorem 3.1 the global error behaves for $t = nh$ like
$$y_n - y(t) = \mathcal{O}(th^p) + \mathcal{O}(t^2 h^q),$$
and the action variables like
$$I(y_n) - I(y_0) = \mathcal{O}(h^p) + \mathcal{O}(th^q).$$
Remark. Methods satisfying the assumptions of this exercise are called *pseudo-symplectic* of order (p, q) (Aubry & Chartier 1998). Pseudo-symplectic methods behave like symplectic methods on time intervals of length $\mathcal{O}(h^{p-q})$.
10. Using the theory of B-series, in particular Theorem VI.7.4, derive the conditions for the coefficients of a Runge–Kutta method such that it is pseudo-symplectic of order $p(q)$. Prove that there exist explicit, pseudo-symplectic Runge–Kutta methods of order $(2, 4)$ with 3 stages.

Chapter XI. Reversible Perturbation Theory and Symmetric Integrators

> There is a very close similarity between the behaviour of solutions of reversible systems and that of Hamiltonian ones.
>
> (M.B. Sevryuk 1986, p. 3)

Numerical experiments indicate that symmetric methods applied to integrable and near-integrable reversible systems share similar properties to symplectic methods applied to (near-)integrable Hamiltonian systems: linear error growth, long-time near-conservation of first integrals, existence of invariant tori. The present chapter gives a theoretical explanation of the good long-time behaviour of symmetric methods. The results and techniques are largely analogous to those of the previous chapter – the extent of the analogy may indeed be seen as the most surprising feature of this chapter.

XI.1 Integrable Reversible Systems

We consider a system of differential equations on a domain of $\mathbb{R}^m \times \mathbb{R}^n$,

$$\begin{aligned} \dot{u} &= f(u,v) \\ \dot{v} &= g(u,v)\ , \end{aligned} \tag{1.1}$$

which is *reversible* with respect to the involution $(u,v) \mapsto (u,-v)$: for all (u,v),

$$\begin{aligned} f(u,-v) &= -f(u,v) \\ g(u,-v) &= \quad g(u,v)\ . \end{aligned} \tag{1.2}$$

From Sect. V.1 we recall that the time-t flow φ_t of a reversible system is a *reversible map*:

$$\varphi_t(u,v) = (\widehat{u},\widehat{v}) \qquad \text{implies} \qquad \varphi_t^{-1}(u,-v) = (\widehat{u},-\widehat{v})\ .$$

A coordinate transform $u = \mu(x,y)$, $v = \nu(x,y)$ is said to *preserve reversibility* if the relations

$$\begin{aligned} \mu(x,-y) &= \quad \mu(x,y) \\ \nu(x,-y) &= -\,\nu(x,y) \end{aligned} \tag{1.3}$$

hold for all (x,y). This implies that every reversible system (1.1) written in the new variables (x,y) is again reversible, and that every reversible map $(u,v) \mapsto (\widehat{u},\widehat{v})$

expressed in the variables (x,y) again becomes a reversible map $(x,y) \mapsto (\widehat{x},\widehat{y})$. Conversely, (1.3) is necessary for these properties.

For Hamiltonian systems, complete integrability is tied to the existence of a symplectic transformation to action-angle variables; see Sect. X.1. For reversible systems, we take the existence of a reversibility-preserving transformation to such variables as the definition of integrability.

Definition 1.1. The system (1.1) is called an *integrable reversible system* if, for every point $(u_0,v_0) \in \mathbb{R}^m \times \mathbb{R}^n$ in the domain of (f,g), there exist a function $\omega = (\omega_1,\dots,\omega_n) : D \to \mathbb{R}^n$ and a diffeomorphism

$$\psi = (\mu,\nu) : D \times \mathbb{T}^n \to U \subset \mathbb{R}^m \times \mathbb{R}^n : (a,\theta) \mapsto (u,v)$$

(with D and U open sets in $\mathbb{R}^m$ and $\mathbb{R}^m \times \mathbb{R}^n$, respectively, and $(u_0,v_0) \in U$), which preserves reversibility and transforms the system (1.1) to the form

$$\begin{aligned} \dot{a} &= 0 \\ \dot{\theta} &= \omega(a)\ . \end{aligned} \tag{1.4}$$

We speak of a *real-analytic integrable reversible system* if all the functions appearing in the above definition are real-analytic.

Example 1.2 (Motion in a Central Field). In Examples X.1.2 and X.1.10 we constructed action-angle variables via a series of transformations

$$\begin{pmatrix} q_1,p_2 \\ p_1,q_2 \end{pmatrix} \overset{\text{(X.1.5)}}{\longrightarrow} \begin{pmatrix} r,p_\varphi \\ \varphi,p_r \end{pmatrix} \overset{\text{(X.1.6)}}{\longrightarrow} \begin{pmatrix} H,L \\ y_1,y_2 \end{pmatrix} \overset{\text{(X.1.15)}}{\longrightarrow} \begin{pmatrix} H,L \\ \theta_1,\theta_2 \end{pmatrix}\ .$$

It is easily verified that all these transformations preserve reversibility. They transform the reversible system

$$\begin{aligned} \dot{q}_1 &= p_1, \quad \dot{p}_2 = -q_2 V'(r)/r \\ \dot{q}_2 &= p_2, \quad \dot{p}_1 = -q_1 V'(r)/r \end{aligned} \tag{1.5}$$

(with $r = \sqrt{q_1^2+q_2^2}$) to the form

$$\begin{aligned} &\dot{H} = 0, \quad \dot{L} = 0 \\ &\dot{\theta}_1 = \frac{2\pi}{T}\,, \quad \dot{\theta}_2 = \frac{\Phi}{T} \end{aligned} \tag{1.6}$$

with $T = T(H,L)$ and $\Phi = \Phi(H,L)$ given by (X.1.12) and (X.1.13).

As the following result shows, it is not incidental that the above transformations preserve reversibility.

Theorem 1.3. *In the situation of the Arnold–Liouville theorem, Theorem X.1.6, let the first integrals* $F_1, \ldots, F_d$ *of the completely integrable Hamiltonian system be such that all* F_i *are even functions of the second half of the arguments:*

$$F_i(u, v) = F_i(u, -v) \qquad (i = 1, \ldots, d) . \tag{1.7}$$

Suppose that $\partial F_1/\partial u, \ldots, \partial F_d/\partial u$ *are linearly independent everywhere (on* $\bigcup\{M_x : x \in B\}$*) except possibly on a set that has no interior points. Further, assume that for every* $x \in B$ *there exists* u *such that* $(u, 0) \in M_x$*. Then, the transformation* $\psi : (a, \theta) \mapsto (u, v)$ *to action-angle variables as given by Theorem X.1.6 preserves reversibility.*

Proof. The result follows by tracing the proofs of Lemma X.1.1, Theorem X.1.4 and Theorem X.1.6.

(a) For F_i satisfying (1.7) and at points where the Jacobian matrix $\partial F/\partial u$ is invertible, the construction of the local symplectic transformation $\ell = (F_1, \ldots, F_d, G_1, \ldots, G_d) : (u, v) \mapsto (x, y)$ shows that the generating function $S(x, v)$ becomes odd in v when the integration constant is chosen such that $S(x, 0) = 0$. By (X.1.4), this implies that ℓ preserves reversibility. A continuity argument used together with the essential uniqueness of the transformation ℓ (see Exercise X.3) does away with the exceptional points where $\partial F/\partial u$ is singular.

(b) In Theorem X.1.4, the construction of $e(x, y) = \varphi_y(\ell^{-1}(x, 0)) =: (u, v)$ is such that

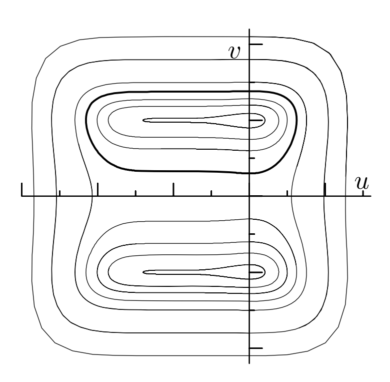
$$e(x, -y) = \varphi_{-y}(\ell^{-1}(x, 0)) = (u, -v) .$$

This holds because by assumption the reference point on M_x can be chosen as $\ell^{-1}(x, 0) = (u_0, 0)$ for some u_0, and because $\varphi_{\pm y}$ is the time ± 1 flow of the Hamiltonian system with Hamiltonian $y_1 F_1 + \ldots + y_d F_d$. Condition (1.7) implies that this is a reversible system, which in turn yields that e preserves reversibility as stated above.

(c) The transformation in the proof of Theorem X.1.6 is of the form $a = w(x)$, $y = W(x)\theta$ (with invertible $W(x) = w'(x)$) and hence preserves reversibility. □

Example 1.4. We now present an example with one degree of freedom where Theorem 1.3 does not apply. In fact, all conditions are satisfied except that for some x there is no u such that $(u, 0) \in M_x$. We consider the Hamiltonian

$$H(u, v) = (v^2 - 1)^2 + \int_0^u s(s + 1)^4 \, ds.$$



Its level sets are shown in the picture to the right. For energy values such that the level curve does not intersect the u-axis, Theorem 1.3 does not apply even though $H(u, v)$ satisfies (1.7). For these energy values the system is an integrable Hamiltonian system, but not an integrable reversible system .

Example 1.5 (Motion in a Central Field, Continued). All the assumptions of Theorem 1.3 are satisfied for $F_1 = H$, $F_2 = L = p_1q_2 - p_2q_1$ if we take the symplectic coordinates $u = (q_1, p_2)$ and $v = (-p_1, q_2)$.

The condition (1.7) is also satisfied with $F_1 = H$, $F_2 = L^2$ ($L \neq 0$ as always) for the choices $u = (p_1, p_2)$ and $v = (q_1, q_2)$, or $u = (q_1, q_2)$ and $v = (-p_1, -p_2)$. However, in these situations, Theorem 1.3 cannot be applied, because there does not exist u such that $(u, 0) \in M_x$.

Example 1.6 (Toda Lattice). Consider the Toda lattice of Sect. X.1.5. The eigenvalues of the matrix L are first integrals in involution. With the symplectic coordinates $(u, v) = (q, -p)$ the Hamiltonian system corresponding to (X.1.17) satisfies the reversibility conditions (1.2). However, since $v_1 + \ldots + v_n$ is a first integral of this system, it is not possible to connect (u, v) with $(u, -v)$ on a level set M_x, and Theorem 1.3 cannot be applied.

Fortunately, as can be seen in Fig. 1.1, the Toda lattice contains many more symmetries. With periodic boundary conditions it is, for example, ρ-reversible (i.e., $\rho f(y) = -f(\rho y)$, $y = (p, q)^T$, see the discussion in Chap. V) with

$$\rho = \begin{pmatrix} S & 0 \\ 0 & -S \end{pmatrix} \qquad S = \begin{pmatrix} & & 1 \\ & 1 & \\ 1 & & \end{pmatrix},$$

where S inverses the components of a vector. To bring the system to the form (1.1) with a vector field satisfying (1.2), we transform S (and hence ρ) to diagonal form and collect the variables corresponding to the eigenvalues $+1$ and -1 in u and v, respectively (see Exercise 1). This gives the (symplectic) coordinates

$$\begin{aligned} u_k &= \frac{1}{\sqrt{2}}\Big(p_k + p_{n-k+1}\Big), & u_{n-k+1} &= \frac{1}{\sqrt{2}}\Big(q_k - q_{n-k+1}\Big), \\ v_k &= \frac{1}{\sqrt{2}}\Big(q_k + q_{n-k+1}\Big), & v_{n-k+1} &= \frac{1}{\sqrt{2}}\Big(p_{n-k+1} - p_k\Big), \end{aligned} \tag{1.8}$$

for $k = 1, \ldots, n/2$ (if n is even; for odd $n = 2\ell + 1$, (1.8) holds for $k = 1, \ldots, \ell$ and in addition we have $u_{\ell+1} = p_{\ell+1}$ and $v_{\ell+1} = q_{\ell+1}$).

In the following we restrict our considerations to the case $n = 3$, and we show that all assumptions of Theorem 1.3 are satisfied, so that we have an integrable reversible system. For $n = 3$, the new variables are

$$\begin{aligned} u_1 &= \frac{1}{\sqrt{2}}\Big(p_1 + p_3\Big), & u_2 &= p_2, & u_3 &= \frac{1}{\sqrt{2}}\Big(q_1 - q_3\Big), \\ v_1 &= \frac{1}{\sqrt{2}}\Big(q_1 + q_3\Big), & v_2 &= q_2, & v_3 &= \frac{1}{\sqrt{2}}\Big(p_3 - p_1\Big), \end{aligned}$$

and the expressions a_k and b_k of Sect. X.1.5 become

$$\begin{aligned} a_1 &= -\frac{1}{2\sqrt{2}}\Big(u_1 - v_3\Big), & b_1 &= \frac{1}{2}\exp\Big(\frac{1}{2}\Big(\frac{1}{\sqrt{2}}(v_1 + u_3) - v_2\Big)\Big), \\ a_2 &= -\frac{1}{2}\,u_2, & b_2 &= \frac{1}{2}\exp\Big(\frac{1}{2}\Big(v_2 - \frac{1}{\sqrt{2}}(v_1 - u_3)\Big)\Big), \\ a_3 &= -\frac{1}{2\sqrt{2}}\Big(u_1 + v_3\Big), & b_3 &= \frac{1}{2}\exp\Big(\frac{1}{\sqrt{2}}\,u_3\Big). \end{aligned}$$

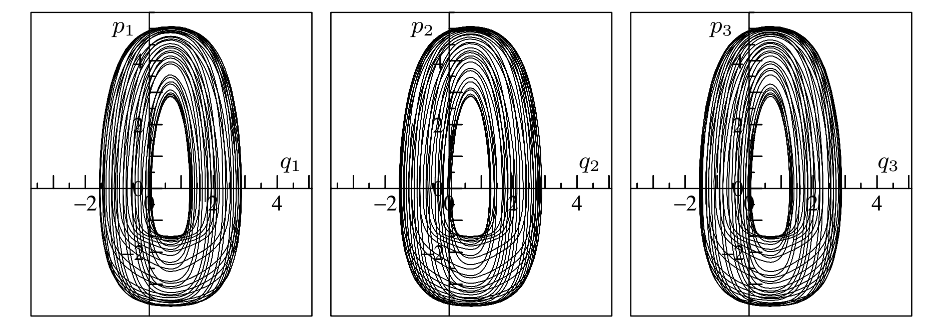


Fig. 1.1. Three projections of the solution of the Toda lattice equations ($n = 3$) with initial values as in Fig. X.1.3

One sees that $b_1^2 + b_2^2$ and $a_1 b_2^2 + a_3 b_1^2$ are even functions of v, so that all coefficients of the characteristic polynomial of the matrix L

$$\begin{aligned}\chi(\lambda) &= -\lambda^3 + (a_1 + a_2 + a_3)\lambda^2 - (a_1a_2 + a_2a_3 + a_3a_1 - b_1^2 - b_2^2 - b_3^2)\lambda + \\ &\quad (a_1a_2a_3 - a_1b_2^2 - a_2b_3^2 - a_3b_1^2 + 2b_1b_2b_3)\ .\end{aligned}$$

are even in v. This implies that also the eigenvalues of L are even functions of v, so that (1.7) is satisfied.

It remains to prove that for fixed x, i.e., for given real eigenvalues of L, the point (u_0, v_0) corresponding to $p(0), q(0)$ can be connected with an element of the form $(u, 0) \in \mathbb{R}^6$ without leaving the level set M_x. Equivalently, we have to find such a path for which the corresponding coefficients of the characteristic polynomial $\chi(\lambda)$ take given values. For given $v(t)$ this yields a system of three nonlinear equations for $u(t) \in \mathbb{R}^3$. For the eigenvalues corresponding to the initial values $p(0), q(0)$ used in Fig. X.1.3, we put $v(t) = v_0 t$ for $1 \geq t \geq 0$ and we check numerically with a path-following algorithm that such a connection is possible.

Example 1.7 (Rigid Body Equations on the Unit Sphere). We reconsider an example that has accompanied us all the way through Chapters IV, V, and VII.5: the rigid body equations (IV.1.4), here considered as differential equations on the unit sphere. We assume $I_3 < I_1, I_2$ for the inertia, which implies that any solution starting with $y_3(0) > 0$ will have $y_3(t) > 0$ for all t. We consider the equations in the neighbourhood of such a solution. We can then choose $u = y_1$, $v = y_2$ as coordinates on the upper half-sphere $\{y_1^2 + y_2^2 + y_3^2 = 1,\ y_3 > 0\}$. This gives the reversible system

$$\begin{aligned}\dot u &= a_1 v\sqrt{1-u^2-v^2} \\ \dot v &= a_2 u\sqrt{1-u^2-v^2}\end{aligned} \tag{1.9}$$

with $a_1 = (I_2 - I_3)/I_2I_3 > 0$ and $a_2 = (I_3 - I_1)/I_3I_1 < 0$, which has $H = u^2/I_1 + v^2/I_2 + (1 - u^2 - v^2)/I_3 = a_2u^2 - a_1v^2 + I_3^{-1}$ as an invariant. We introduce polar coordinates $u = r\cos\varphi$, $v = r\sin\varphi$ and express r as a function of H and φ:

$$r = \sqrt{\frac{I_3^{-1} - H}{a_1 \sin^2 \varphi - a_2 \cos^2 \varphi}} .$$

This leaves us with differential equations

$$\dot{H} = 0, \qquad \dot{\varphi} = \gamma(H, \varphi),$$

where γ is even in φ and has no zeros. The time needed to run through an angle φ is

$$\tau(H, \varphi) = \int_0^\varphi \frac{1}{\gamma(H, \phi)} \, d\phi \, , \qquad \text{and} \qquad \omega(H) = \frac{2\pi}{\tau(H, 2\pi)}$$

is the frequency. With $\theta = \omega(H)\tau(H, \varphi)$ we then have

$$\dot{H} = 0, \qquad \dot{\theta} = \omega(H) \, .$$

The transformation from (u, v) in the open unit disc (except the origin) to $(H, \theta) \in (0, I_3^{-1}) \times \mathbb{T}$ is a diffeomorphism that preserves reversibility. This shows that the rigid body equations (1.9) are an integrable reversible system.

Example 1.8 (Rigid Body Equations in $\mathbb{R}^3$). We now consider the rigid body equations (IV.1.4) in the ambient space $\mathbb{R}^3$, rather than on the unit sphere. The system then has the invariants $H = y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3$ and $K = y_1^2 + y_2^2 + y_3^2$, and it is reversible with respect to the partition $u = (y_1, y_3)$ and $v = y_2$. In the case $I_3 < I_1, I_2$ we can again restrict our attention to $y_3 > 0$. We then write $y_3 = \sqrt{K - y_1^2 - y_2^2}$ and introduce polar coordinates $y_1 = r \cos \varphi$, $y_2 = r \sin \varphi$. As above, we express r as a function of H, K and φ (this just requires replacing I_3^{-1} with K/I_3 in the above formula for r) and we obtain differential equations

$$\dot{H} = 0, \quad \dot{K} = 0, \quad \dot{\varphi} = \gamma(H, K, \varphi)$$

with γ even in φ and without zeros. In the same way as above, this is transformed to

$$\dot{H} = 0, \quad \dot{K} = 0, \quad \dot{\theta} = \omega(H, K) \, .$$

The transformation $((y_1, y_3), y_2) \mapsto ((H, K), \theta)$ preserves reversibility. The rigid body equations (IV.1.4) are thus an integrable reversible system. Note that this time the dimensions differ.

XI.2 Transformations in Reversible Perturbation Theory

We consider perturbations of an integrable reversible system such that the perturbed system is still reversible. This takes the form

$$\begin{aligned} \dot{a} &= \varepsilon r(a,\theta) \\ \dot{\theta} &= \omega(a) + \varepsilon\rho(a,\theta) \end{aligned} \tag{2.1}$$

where ε is a small parameter, and r is an odd function of θ and ρ is an even function of θ:

$$\begin{aligned} r(a,-\theta) &= -r(a,\theta) \\ \rho(a,-\theta) &= \rho(a,\theta) \;. \end{aligned} \tag{2.2}$$

Similar to Sect. X.2 for Hamiltonian perturbation theory, we study coordinate transformations that change (2.1) to reversible systems which – in various ways – look closer to an integrable system in action-angle variables than (2.1).

XI.2.1 The Basic Scheme of Reversible Perturbation Theory

We look for a transformation between neighbourhoods of $\{a_0\} \times \mathbb{T}^n$,

$$\begin{aligned} a &= b + \varepsilon s(b,\varphi) \\ \theta &= \varphi + \varepsilon\sigma(b,\varphi) \;, \end{aligned} \tag{2.3}$$

which preserves reversibility and hence has s even in φ and σ odd in φ, such that the transformed system is of the form

$$\begin{aligned} \dot{b} &= \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon\mu(b) + \mathcal{O}(\varepsilon^2) \;. \end{aligned} \tag{2.4}$$

Inserting (2.3) into (2.1) gives the system

$$\left\{ \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \varepsilon \begin{pmatrix} \partial s/\partial b & \partial s/\partial\varphi \\ \partial\sigma/\partial b & \partial\sigma/\partial\varphi \end{pmatrix} \right\} \begin{pmatrix} \dot{b} \\ \dot{\varphi} \end{pmatrix} = \begin{pmatrix} \varepsilon r(a,\theta) \\ \omega(a) + \varepsilon\rho(a,\theta) \end{pmatrix}$$

with (a,θ) from (2.3). Inverting the matrix on the left-hand side and expanding in powers of ε, it is seen that (2.4) requires that s, σ satisfy the equations

$$\frac{\partial s}{\partial\varphi}(b,\varphi)\,\omega(b) = r(b,\varphi) \tag{2.5}$$

$$\frac{\partial\sigma}{\partial\varphi}(b,\varphi)\,\omega(b) = \rho(b,\varphi) + \omega'(b)\,s(b,\varphi) - \mu(b) \;. \tag{2.6}$$

A necessary condition for the solvability of (2.5) is that the angular average of r vanishes:

$$\overline{r}(b) = 0 \;, \qquad \text{where} \quad \overline{r}(b) = \frac{1}{(2\pi)^n}\int_{\mathbb{T}^n} r(b,\varphi)\,d\varphi \;. \tag{2.7}$$

In the Hamiltonian case this condition was satisfied because r was a gradient with respect to φ. Here, in the reversible case, this is satisfied because r is an odd function of φ.

If (2.7) holds, then (2.5) can be solved by Fourier series expansion in the same way as we solved (X.2.2), provided that the frequencies $\omega_1(b), \ldots, \omega_n(b)$ are non-resonant. Of course, there is again the same problem of small denominators as in the Hamiltonian case. Equations (2.6) are solved in the same way as (2.5), upon setting

$$\mu(b) \;=\; \overline{\rho}(b) + \omega'(b)\,\overline{s}(b)\,. \tag{2.8}$$

Since r is odd in φ, the solution s of (2.5) becomes even in φ. It is determined uniquely only up to a constant: we are still free to choose the angular average $\overline{s}(b)$. If $\omega'(b)$ has rank n, we may actually choose $\overline{s}(b)$ such that $\mu(b) = 0$ results from (2.8). Since the right-hand side of (2.6) is even in φ, the solution σ of (2.6) becomes odd in φ if we choose $\overline{\sigma}(b) = 0$.

XI.2.2 Reversible Perturbation Series

The above construction extends to arbitrary finite order in ε. The transformation is now sought for in the form

$$a \;=\; b + \varepsilon s_1(b,\varphi) + \varepsilon^2 s_2(b,\varphi) + \ldots + \varepsilon^{N-1} s_{N-1}(b,\varphi) \tag{2.9}$$

$$\theta \;=\; \varphi + \varepsilon \sigma_1(b,\varphi) + \varepsilon^2 \sigma_2(b,\varphi) + \ldots + \varepsilon^{N-1} \sigma_{N-1}(b,\varphi) \tag{2.10}$$

with s_j even in φ and σ_j odd in φ to preserve reversibility. This transformation is to be chosen such that the system in the new variables is of the form

$$\begin{aligned} \dot b &= \varepsilon^N r_N(b,\varphi) \\ \dot\varphi &= \omega_{\varepsilon,N}(b) + \varepsilon^N \rho_N(b,\varphi) \end{aligned}$$

with $\omega_{\varepsilon,N}(b) = \omega(b) + \varepsilon\mu_1(b) + \ldots + \varepsilon^{N-1}\mu_{N-1}(b)$, and with $r_N(b,\varphi)$ odd in φ and $\rho_N(b,\varphi)$ even in φ, and with all these functions bounded independently of ε.

Inserting the transformation into (2.1) and expanding in powers of ε, it is seen that the functions s_j and σ_j must satisfy equations of the form of (2.5), (2.6):

$$\frac{\partial s_j}{\partial\varphi}(b,\varphi)\,\omega(b) \;=\; p_j(b,\varphi) \tag{2.11}$$

$$\frac{\partial \sigma_j}{\partial\varphi}(b,\varphi)\,\omega(b) \;=\; \pi_j(b,\varphi) + \omega'(b)\,s_j(b,\varphi) - \mu_j(b) \tag{2.12}$$

where p_j, π_j are given by expressions that depend linearly on higher-order derivatives of r, ρ and polynomially on the functions s_i, σ_i with $i < j$ and on their first-order derivatives. Using the rules

$$\begin{pmatrix} \text{even} & \text{odd} \\ \text{odd} & \text{even} \end{pmatrix} \begin{pmatrix} \text{odd} \\ \text{even} \end{pmatrix} = \begin{pmatrix} \text{odd} \\ \text{even} \end{pmatrix}$$

and

$$\frac{\partial\,\text{even}}{\partial\varphi} = \text{odd}\ , \qquad \frac{\partial\,\text{odd}}{\partial\varphi} = \text{even}\ ,$$

it is found that p_j is odd in φ and π_j is even in φ for all j. For non-resonant frequencies $\omega(b)$, the equations (2.11), (2.12) can therefore be solved with s_j even in φ, σ_j odd in φ. If $\omega'(b)$ is invertible, we can obtain $\mu_j(b) = 0$ for all j.

Beyond these formal calculations, there is the following reversible analogue of Lemma X.2.1 in the Hamiltonian case. This result is obtained by the same "ultraviolet cut-off" argument as the earlier result.

Lemma 2.1. *Let the right-hand side functions of (2.1) be real-analytic in a neighbourhood of $\{b^*\} \times \mathbb{T}^n$ and satisfy (2.2). Suppose that $\omega(b^*)$ satisfies the diophantine condition (X.2.4). For any fixed $N \geq 2$, there are positive constants ε_0, c, C such that the following holds for $\varepsilon \leq \varepsilon_0$: there exists a real-analytic reversibility-preserving change of coordinates $(a,\theta) \mapsto (b,\varphi)$ such that every solution $(b(t),\varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \leq c|\log\varepsilon|^{-\nu-1}$, satisfies*

$$\begin{aligned} &\|b(t) - b(0)\| \leq C\,t\,\varepsilon^N \qquad \textit{for} \ \ t \leq \varepsilon^{-N+1}\ , \\ &\|\varphi(t) - \omega_{\varepsilon,N}(b(0))t - \varphi(0)\| \leq C\,(t^2 + t|\log\varepsilon|^{\nu+1})\,\varepsilon^N \ \ \textit{for} \ \ t^2 \leq \varepsilon^{-N+1}. \end{aligned}$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$-close to the identity: $\|(a,\theta) - (b,\varphi)\| \leq C\varepsilon$ holds for (a,θ) and (b,φ) related by the above coordinate transform, for $\|b - b^*\| \leq c|\log\varepsilon|^{-\nu-1}$ and for φ in an ε-independent complex neighbourhood of $\mathbb{T}^n$.

The constants ε_0, c, C depend on N, n, γ, ν and on bounds of ω, r, ρ on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^n$. □

The equations determining the coefficient functions of the perturbation series are of the form to which Lemma X.4.1 applies. Therefore, that lemma is again the tool for estimating the terms in the perturbation series, similar to Sect. X.4.1. This yields a reversible analogue of Theorem X.4.4 showing near-invariance of tori (up to exponentially small terms in a negative power of ε) over time intervals that are exponentially large in a negative power of ε, with the same exponents α, β as in Theorem X.4.4.

XI.2.3 Reversible KAM Theory

For an integrable reversible system, just as for an integrable Hamiltonian system, the phase space is foliated into invariant tori on which the flow is conditionally periodic. We fix one such torus $\{a = a^*,\ \theta \in \mathbb{T}^n\}$ with diophantine frequencies $\omega_1, \ldots, \omega_n$. For convenience we may assume $a^* = 0 \in \mathbb{R}^m$. This torus is invariant under the flow of systems of the form $\dot{a} = \mathcal{O}(\|a\|^2)$, $\dot{\theta} = \omega + \mathcal{O}(\|a\|)$, or written more explicitly,

$$\begin{aligned} \dot{a} &= \tfrac{1}{2} a^T K(a,\theta) a \\ \dot{\theta} &= \omega + M(a,\theta) a\ . \end{aligned} \tag{2.13}$$

Here, $K = [K_1, \ldots, K_m]$ where each $K_i(a,\theta)$ is a symmetric $m \times m$ matrix, and $M(a,\theta)$ is an $n \times m$ matrix. The first equation is to be interpreted as $\dot a_i = \frac{1}{2} a^T K_i(a,\theta) a$ for the components $i = 1, \ldots, m$. Consider now a perturbation of this system:

$$\begin{aligned} \dot a &= \tfrac{1}{2} a^T K(a,\theta) a + \varepsilon r(a,\theta) \\ \dot\theta &= \omega + M(a,\theta) a + \varepsilon \rho(a,\theta)\,. \end{aligned} \tag{2.14}$$

For the reversible case, i.e., for K and r odd in θ and for M and ρ even in θ, we construct a sequence of reversibility-preserving transformations in the spirit of Kolmogorov's transformation of Sect. X.2.3, which transform (2.14) back to the form (2.13) in the new variables, showing the persistence of an invariant torus with frequencies ω_i under small reversible perturbations of the system. This holds again under the diophantine condition (X.2.4) on ω and additionally under the condition that the angular average $\overline{M}_0$ of M at $a = 0$ has rank n. A result of this type – a reversible KAM theorem – was shown by Moser (1973), Chap. V, in a different setting. See also Sevryuk (1986) for further results in that direction.

We look for a transformation of the form

$$\begin{aligned} a &= b + \varepsilon\Big(s(\varphi) + S(\varphi) b\Big) \\ \theta &= \varphi + \varepsilon \sigma(\varphi) \end{aligned} \tag{2.15}$$

with an $m \times m$ matrix $S(\varphi)$. Preserving reversibility requires that s and S are even functions and σ is odd. Higher-order terms in b play no role and are therefore omitted from the beginning. We insert this into (2.14) and obtain

$$\begin{aligned} \dot b &= \frac{1}{2}\, b^T K(b,\varphi) b + \varepsilon \Big\{ r(0,\varphi) - \frac{\partial s}{\partial \varphi}(\varphi)\omega \\ &\quad + \frac{\partial r}{\partial b}(0,\varphi) b - \frac{\partial s}{\partial \varphi}(\varphi) M(0,\varphi) b - \frac{\partial}{\partial \varphi}\Big(S(\varphi) b\Big)\omega + s(\varphi)^T K(0,\varphi) b \Big\} \\ &\quad + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \|b\|^2) \\ \dot\varphi &= \omega + M(b,\varphi) b \\ &\quad + \varepsilon \Big\{ \rho(0,\varphi) - \frac{\partial \sigma}{\partial \varphi}(\varphi)\omega + M(0,\varphi) s(\varphi) \Big\} + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon \|b\|)\,. \end{aligned}$$

We require that the terms in curly brackets vanish. This holds if the following equations are satisfied (the last equation is written component-wise for notational clarity):

$$\begin{aligned} \frac{\partial s}{\partial \varphi}(\varphi)\,\omega &= r(0,\varphi) \\ \frac{\partial \sigma}{\partial \varphi}(\varphi)\,\omega &= \rho(0,\varphi) + M(0,\varphi) s(\varphi) \\ \frac{\partial S_{ij}}{\partial \varphi}(\varphi)\,\omega &= \frac{\partial r_i}{\partial b_j}(\varphi) - \sum_k \frac{\partial s_i}{\partial \varphi_k}(\varphi) M_{kj}(0,\varphi) + \sum_k s_k(\varphi) K_{i,kj}(0,\varphi)\,. \end{aligned} \tag{2.16}$$

Since r is odd in φ, the first equation can be solved for s even in φ, uniquely up to a constant, the angular average $\overline{s}$. Since the angular average of M is assumed to be of full rank n, $\overline{s}$ can be chosen such that the angular average of the right-hand side of the equation for σ becomes zero. Since the right-hand side is even, the equation can then be solved uniquely for an odd σ. The equations for S have an odd right-hand side and can therefore be solved for an even S.

In this way, the perturbation to the form (2.13) is reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. By the same arguments as in the Hamiltonian case (see Sect. X.5), the iteration of this procedure is seen to be convergent. This finally yields a change of coordinates that preserves reversibility and transforms the perturbed system (2.14) back to the form (2.13). We summarize this in the following theorem, which is the reversible analogue of Kolmogorov's Theorem X.5.1.

Theorem 2.2. *Consider a real-analytic reversible system (2.13). Suppose that $\omega \in \mathbb{R}^n$ satisfies the diophantine condition (X.2.4), and that the angular average of $M(0,\cdot)$ is an $n \times m$ matrix of rank n. Let (2.14) be a real-analytic reversible perturbation of the system (2.13). Then, there exists $\varepsilon_0 > 0$ (which depends on the perturbation functions only through a bound of their norms on a complex neighbourhood of $\{0\} \times \mathbb{T}^n$) such that for every ε with $|\varepsilon| \le \varepsilon_0$, there is a real-analytic transformation $\psi_\varepsilon : (b,\varphi) \mapsto (a,\theta)$, $\mathcal{O}(\varepsilon)$ close to the identity and depending analytically on ε, which preserves reversibility and puts the perturbed system back to the form (2.13) in the new variables: $\dot b = \mathcal{O}(\|b\|^2)$, $\dot\varphi = \omega + \mathcal{O}(\|b\|)$. The perturbed system therefore has the invariant torus $\{b = 0,\ \varphi \in \mathbb{T}^n\}$ carrying a quasi-periodic flow with the same frequencies ω as the unperturbed system.* □

XI.2.4 Reversible Birkhoff-Type Normalization

We show that, in the situation of diophantine frequencies ω, there is a reversibility-preserving transformation that takes a reversible system of the form (2.13) to the form

$$\begin{aligned} \dot b &= r_k(b,\varphi) \\ \dot\varphi &= \omega + \zeta_k(b) + \rho_k(b,\varphi) \end{aligned} \qquad \text{with} \qquad r_k, \rho_k = \mathcal{O}(\|b\|^k) \tag{2.17}$$

for arbitrary $k \ge 2$, where $\zeta_k = \overline{\rho}_1 + \ldots + \overline{\rho}_{k-1}$ with the bars denoting angular averages and with $\rho_1(b,\varphi) = M(b,\varphi)b$. This implies again that the invariant torus is "very sticky": $\|b(0)\| \le \delta$ implies $\|b(t)\| \le 2\delta$ for $t \le C_k\delta^{-k+1}$. As in the Hamiltonian case, a suitable choice of k would even yield time intervals exponentially long in a negative power of δ during which solutions stay within twice the initial distance δ.

The transformation to the normal form (2.17) is constructed recursively. Suppose that in some variables (a,θ) we have, for some $k \ge 2$,

$$\begin{aligned} \dot a &= r_{k-1}(a,\theta) \\ \dot\theta &= \omega + \zeta_{k-1}(a) + \rho_{k-1}(a,\theta) \end{aligned} \qquad \text{with} \qquad r_{k-1}, \rho_{k-1} = \mathcal{O}(\|a\|^{k-1})\ .$$

Note, for $k = 2$ we have $r_1 = \mathcal{O}(\|a\|^2)$ by (2.13). We search for a transformation

$$\begin{aligned} a &= b + s(b,\varphi) \\ \theta &= \varphi + \sigma(b,\varphi) \end{aligned} \qquad \text{with} \qquad s,\sigma = \mathcal{O}(\|b\|^{k-1}) \, ,$$

(and $s = \mathcal{O}(\|b\|^2)$ for $k = 2$) that preserves reversibility, i.e., has s even in φ and σ odd in φ, and is such that (2.17) holds. Inserting the transformation into the above differential equation shows that this is indeed achieved if s, σ solve the following system of the form (2.5), (2.6):

$$\begin{aligned} \frac{\partial s}{\partial \varphi}(b,\varphi)\,\omega &= r_{k-1}(b,\varphi) \\ \frac{\partial \sigma}{\partial \varphi}(b,\varphi)\,\omega &= \rho_{k-1}(b,\varphi) + \zeta_{k-1}'(b)\,s(b,\varphi) - \mu_k(b) \, . \end{aligned}$$

Choosing $\overline{s}(b) = 0$ leads to $\mu_k = \overline{\rho}_{k-1}$ and gives (2.17) with $\zeta_k = \zeta_{k-1} + \overline{\rho}_{k-1}$.

XI.3 Linear Error Growth and Near-Preservation of First Integrals

We now study the error behaviour of reversible methods applied to integrable reversible systems. Recall from Theorem V.1.5 that symmetric methods are reversible under the compatibility condition (V.1.4). We give an analogue of Theorem X.3.1 on the error behaviour of symplectic methods applied to integrable Hamiltonian systems. We consider an integrable reversible system (1.1) (usually not given in action-angle variables) and let $(u,v) = \psi(a,\theta)$ be the reversibility-preserving transformation to action-angle variables. The inverse transformation is denoted as

$$(a,\theta) = (I(u,v), \Theta(u,v)) \, .$$

The following is the reversible analogue of Theorem X.3.1.

Theorem 3.1. *Consider applying a reversible numerical integrator of order p to the integrable reversible system (1.1) with real-analytic right-hand side. Suppose that $\omega(a^*)$ satisfies the diophantine condition (X.2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \le h_0$: every numerical solution starting with $\|I(u_0,v_0) - a^*\| \le c\,|\log h|^{-\nu-1}$ satisfies*

$$\begin{aligned} \|(u_n,v_n) - (u(t),v(t))\| &\le C\,t\,h^p \\ \|I(u_n,v_n) - I(u_0,v_0)\| &\le C\,h^p \end{aligned} \qquad \text{for } \ t = nh \le h^{-p}. \tag{3.1}$$

The constants h_0, c, C depend on γ, ν of (X.2.4), on the dimensions, on bounds of the real-analytic functions f, g on a complex neighbourhood of the torus $\{(u,v) : I(u,v) = a^*\}$, and on the numerical method.

Proof. The proof of Theorem X.3.1 relied on Theorem IX.3.1 and Lemma X.2.1. Using their reversible analogues Theorem IX.2.3 and Lemma 2.1 with the same arguments gives the above result for the reversible case. □

Remark 3.2. As in the analogous remark for the Hamiltonian case, the error bounds of Theorem 3.1 also hold when the reversible method is applied to a perturbed integrable system with a perturbation parameter ε bounded by a positive power of the step size: $\varepsilon \leq Kh^\alpha$ for some $\alpha > 0$.

We consider the Hamiltonian system of Example 1.4 and apply the symmetric but non-symplectic Lobatto IIIB method with step size $h = 0.01$. In the left picture of Fig. 3.1 we choose the initial value $(u_0, v_0) = (0, 1.5)$ for which the level curve of the Hamiltonian is symmetric with respect to the u-axis and the system is an integrable reversible system. The good conservation of the Hamiltonian is in agreement with Theorem 3.1. In the right picture we choose $(u_0, v_0) = (0, 0.3)$ whose level curve is the fat line in the picture of Example 1.4 which does not intersect the u-axis. Since in this situation we do not have an integrable reversible system, Theorem 3.1 cannot be applied and we cannot expect good energy conservation.

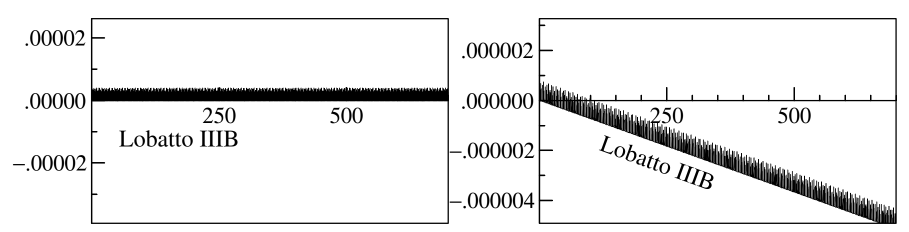


Fig. 3.1. Numerical Hamiltonian of Example 1.4 for two different initial values

For the Toda lattice example, Figures 3.2 and 3.3 illustrate the long-time conservation of the first integrals and the linear error growth, respectively, of the Lobatto IIIB method.

Theorem 3.1 together with Examples 1.7 and 1.8 also explains the good behaviour of symmetric (in fact, reversible) integrators on the rigid body equations which we observed in Chap. V (Figs. V.4.2 and V.4.6).

Variable Step Sizes: Proportional, Reversible Controllers. As a consequence of the backward error analysis of Theorem IX.6.1 the statement (3.1) can be extended straightforwardly to proportional step size controllers as discussed in Sect. VIII.3.1. Under the assumption of Theorem 3.1 with h and h_0 replaced by ε and ε_0 one has

$$\begin{aligned} \|(u_n, v_n) - (u(t_n), v(t_n))\| &\leq C\, t_n\, \varepsilon^p \\ \|I(u_n, v_n) - I(u_0, v_0)\| &\leq C\, \varepsilon^p \end{aligned} \qquad \text{for } t_n \leq \varepsilon^{-p}. \tag{3.2}$$

The grid $\{t_n\}$ is determined by the method and satisfies $t_{n+1} = t_n + \varepsilon s(u_n, v_n, \varepsilon)$.

Variable Step Sizes: Integrating, Reversible Controllers. We apply the backward error analysis of Theorem IX.6.2. The modified equation (IX.6.14) reduces to

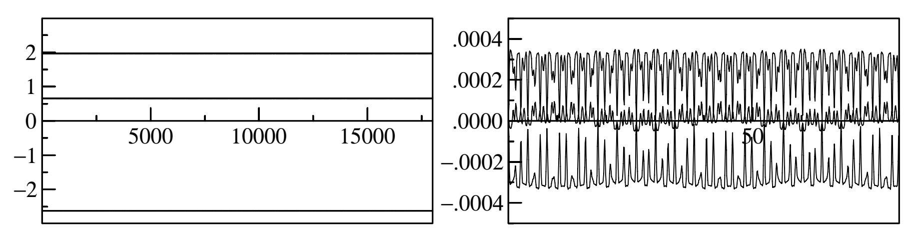


Fig. 3.2. Numerically obtained eigenvalues (left picture) and errors in the eigenvalues (right picture) of the 3-stage Lobatto IIIB scheme (step size $h = 0.1$) applied to the Toda lattice with the data of Sect. X.1.5

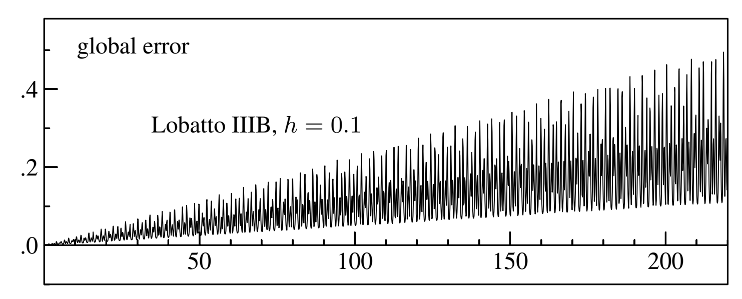


Fig. 3.3. Euclidean norm of the global error for the 3-stage Lobatto IIIB scheme (step size $h = 0.1$) applied to the Toda lattice with $n = 3$ and initial values as in Fig. 3.2

$$\dot{y} = f(y), \qquad \dot{z} = z\, G(y) \tag{3.3}$$

for $\varepsilon = 0$. Since $G(y) = -\big(\sigma(y)\big)^{-1}\nabla\sigma(y)^T f(y)$ with an analytic step size function $\sigma(y)$, the function $(y, z) \mapsto z\sigma(y)$ is a first integral of (3.3). Suppose now that $\dot{y} = f(y)$ is the integrable reversible system (1.1). This means that there exists a reversibility preserving diffeomorphism $y = \psi(a, \theta)$ transforming the system to action-angle variables. The diffeomorphism

$$\begin{pmatrix} y \\ z \end{pmatrix} = \widehat{\psi}(a, A, \theta) = \begin{pmatrix} \psi(a, \theta) \\ A/\sigma\big(\psi(a, \theta)\big) \end{pmatrix}$$

is then also reversibility preserving if $\sigma(u, -v) = \sigma(u, v)$, and it transforms (3.3) to

$$\dot{a} = 0, \qquad \dot{A} = 0, \qquad \dot{\theta} = \omega(a).$$

If the basic method of the algorithm (IX.6.9) is reversible and if $\sigma(u, -v) = \sigma(u, v)$ holds, the modified equation (IX.6.14) is a reversible perturbation of (3.3). Consequently, Theorem 3.1 yields the statement (3.2) also for integrating step size controllers. Since $A := z\,\sigma(u, v)$ is an action variable, we have in addition that

$$|z_n\sigma(u_n, v_n) - z_0\sigma(u_0, v_0)| \le C\varepsilon^2$$

for $t_n \le \varepsilon^{-p}$. Notice that the transformation (2.9) is $\mathcal{O}(\varepsilon^p)$-close to the identity for the variables a and θ, but only $\mathcal{O}(\varepsilon^2)$-close for A. This result proves that the integrating step size controller is as robust as the proportional controller. It also explains the excellent long-time behaviour observed in Figs. VIII.3.2 and VIII.3.3.

XI.4 Invariant Tori under Reversible Discretization

In this section we study the question as to how invariant tori of reversible systems are preserved under discretization of the system by reversible numerical methods. We give reversible analogues of Theorems X.5.3 and X.6.1.

XI.4.1 Near-Invariant Tori over Exponentially Long Times

We consider a reversible system (1.1) which in suitable coordinates takes the perturbed form (2.14). Under the conditions of the reversible KAM theorem, Theorem 2.2, this system has an invariant torus carrying a quasi-periodic flow with frequencies ω for sufficiently small ε. Consider now a reversible numerical integrator applied to this system. By the same arguments as in Sect. X.5.2, using the reversible KAM theorem 2.2 in place of Kolmogorov's Theorem X.5.1, we obtain the following analogue of Theorem X.5.3, which states the existence of a torus such that numerical solutions starting on this torus remain exponentially close to a quasi-periodic flow on that torus over exponentially long times in $1/h$.

Theorem 4.1. *In the above situation, for a reversible numerical method of order p used with sufficiently small step size h, there is a modified reversible system with an invariant torus $\widetilde{\mathcal{T}}_\omega$ carrying a quasi-periodic flow with frequencies ω, $\mathcal{O}(h^p)$ close to the invariant torus $\mathcal{T}_\omega$ of the original reversible system, such that the difference between any numerical solution (u_n, v_n) starting on the torus $\widetilde{\mathcal{T}}_\omega$ and the solution $(\widetilde{u}(t), \widetilde{v}(t))$ of the modified Hamiltonian system with the same starting values remains exponentially small in $1/h$ over exponentially long times:*

$$\|(u_n, v_n) - (\widetilde{u}(t), \widetilde{v}(t))\| \leq Ce^{-\kappa/h} \quad \textit{for} \quad t = nh \leq e^{\kappa/h} .$$

The constants C and κ are independent of h, ε (for h, ε sufficiently small) and of the initial value $(u_0, v_0) \in \widetilde{\mathcal{T}}_\omega$. □

The case of initial values lying close to, but not on $\widetilde{\mathcal{T}}_\omega$, can again be treated by a reversible analogue of Theorem X.4.7.

XI.4.2 A KAM Theorem for Reversible Near-Identity Maps

To obtain truly invariant tori, we need a discrete analogue of the reversible KAM theorem, which is derived in this subsection. This result can also be viewed as the reversible analogue of Theorem X.6.1. It establishes the existence of invariant tori of reversible integrators, but as in the symplectic case, only for a Cantor set of non-resonant step sizes.

A map $\Phi : (a, \theta) \mapsto (\widehat{a}, \widehat{\theta})$ has the invariant torus $\{a = 0,\ \theta \in \mathbb{T}^n\}$, and reduces on this torus to rotation by $h\omega$ (h a real parameter and $\omega \in \mathbb{R}^n$), when it is of the form (cf. (2.13))

$$\begin{aligned}\widehat{a} &= a + \tfrac{1}{2}ha^T K(a,\theta)a \\ \widehat{\theta} &= \theta + h\omega + hM(a,\theta)a\,.\end{aligned} \tag{4.1}$$

Here, $K = [K_1, \ldots, K_m]$ where each $K_i(a,\theta)$ is a symmetric $m \times m$ matrix, and $M(a,\theta)$ is an $n \times m$ matrix. The expression in the first equation is again to be interpreted as $a^T K_i(a,\theta)a$ for the components $i = 1, \ldots, m$.

A necessary condition for the above map Φ to be *reversible* with respect to the involution $(a,\theta) \mapsto (a,-\theta)$, cf. Definition V.1.2, is seen to be

$$\begin{aligned}K(0,-\theta) &= -K(0,\theta - h\omega) \\ M(0,-\theta) &= M(0,\theta - h\omega)\,.\end{aligned} \tag{4.2}$$

Consider now a perturbed map

$$\begin{aligned}\widehat{a} &= a + \tfrac{1}{2}ha^T K(a,\theta)a + h\,\varepsilon r(a,\theta) \\ \widehat{\theta} &= \theta + h\omega + hM(a,\theta)a + h\,\varepsilon\rho(a,\theta)\end{aligned} \tag{4.3}$$

where r and ρ, which like K and M are assumed real-analytic, might depend analytically also on h and ε. Reversibility of this map implies, by direct computation, that in addition to (4.2), the following equations are satisfied up to an error $\mathcal{O}(h\varepsilon)$:

$$\begin{aligned}r(0,-\theta) &= -\,r(0,\theta - h\omega) \\ \frac{\partial r}{\partial a}(0,-\theta) &= -\frac{\partial r}{\partial a}(0,\theta) \\ \rho(0,-\theta) &= \rho(0,\theta - h\omega) - hM(0,\theta - h\omega)r(0,\theta - h\omega)\,.\end{aligned} \tag{4.4}$$

Similar to Sect. XI.2.3, we construct a reversibility-preserving near-identity transformation of coordinates $(a,\theta) \mapsto (b,\varphi)$ such that the above map $\Phi_{h,\varepsilon}$ in the new variables is of the form (4.3) with the perturbation terms reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. Similar to Sect. X.6.1, this is possible if $h\omega$ satisfies the diophantine condition (X.6.3) and if the angular average $\overline{M}_0$ of $M(0,\cdot)$ has rank n.

We look for the transformation in the form (2.15). The functions defining this transformation must satisfy the following equations, cf. (2.16):

$$\begin{aligned}\frac{s(\varphi + h\omega) - s(\varphi)}{h} &= r(0,\varphi) \\ \frac{\sigma(\varphi + h\omega) - \sigma(\varphi)}{h} &= \rho(0,\varphi) + M(0,\varphi)s(\varphi) \\ \frac{S_{ij}(\varphi + h\omega) - S_{ij}(\varphi)}{h} &= \frac{\partial r_i}{\partial b_j}(\varphi) - \sum_k \frac{\partial s_i}{\partial \varphi_k}(\varphi)M_{kj}(0,\varphi) \\ &\qquad + \sum_k s_k(\varphi)K_{i,kj}(0,\varphi)\,.\end{aligned} \tag{4.5}$$

Under the conditions (X.6.3), (X.6.4) these equations can be solved by Fourier expansion, in the same way as the analogous equations in Sections X.6.1 and XI.2.3, and the map in the variables (b,φ) becomes of the form

$$\begin{aligned} \widehat{b} &= b + \tfrac{1}{2} h b^T K(b,\varphi) b + \mathcal{O}(h\varepsilon \|b\|^2) + \mathcal{O}(h\varepsilon^2) \\ \widehat{\varphi} &= \varphi + h\omega + hM(b,\varphi) b + \mathcal{O}(h\varepsilon \|b\|) + \mathcal{O}(h\varepsilon^2)\ . \end{aligned} \tag{4.6}$$

We still need to know that the change of variables $(a,\theta) \mapsto (b,\varphi)$ preserves reversibility, i.e., that s and S are even functions of φ and σ is an odd function of φ. This is indeed a consequence of (4.2) and (4.4). (We may modify r and ρ such that (4.4) holds exactly, at the expense of introducing additional $\mathcal{O}(h^2\varepsilon^2)$ perturbations in (4.3).) Let us show this property for s. The Fourier coefficients s_k of s must satisfy

$$\frac{e^{ik\cdot h\omega} - 1}{h}\, s_k = r_k\ .$$

Since (4.4) implies $r_{-k} = -r_k e^{-ik\cdot h\omega}$ for all k, it follows that $s_{-k} = s_k$, and hence s is an even function of φ. Similarly it is shown that S is even and σ is odd.

In summary, we have found a transformation $\mathcal{O}(\varepsilon)$ close to the identity, which transforms the reversible map (4.3) to a reversible map (4.6), thus reducing the perturbation terms from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. The iteration of this procedure can again be shown to be convergent. This finally yields a transformation to coordinates in terms of which the perturbed map is back in the form (2.13). In this way we obtain the following discrete analogue of Theorem 2.2 or reversible analogue of Theorem X.6.1.

Theorem 4.2. *Consider a real-analytic reversible map $\Phi_{h,\varepsilon}$ of the form (4.3), defined on a neighbourhood of $\{0\} \times \mathbb{T}^n$, with $0 \in \mathbb{R}^m$. Suppose that $h\omega$ satisfies the diophantine condition (X.6.3), and that the angular average of $M(0,\cdot)$ has rank n. Then, there exists $\varepsilon_0 > 0$ such that for every ε with $|\varepsilon| < \varepsilon_0$, there is a real-analytic transformation $\psi_{h,\varepsilon} : (b,\varphi) \mapsto (a,\theta)$, which preserves reversibility and is $\mathcal{O}(\varepsilon)$ close to the identity uniformly in h satisfying (X.6.3) and is analytic in ε, such that $\psi_{h,\varepsilon}^{-1} \circ \Phi_{h,\varepsilon} \circ \psi_{h,\varepsilon} : (b,\varphi) \mapsto (\widehat{b},\widehat{\varphi})$ is again of the form (4.1): $\widehat{b} = b + \mathcal{O}(\|b\|^2)$, $\widehat{\varphi} = \varphi + h\omega + \mathcal{O}(\|b\|)$. The perturbed map $\Phi_{h,\varepsilon}$ therefore has an invariant torus on which it is conjugate to rotation by $h\omega$.* □

As in the analogous situation of Sect. X.6.2, Theorem 4.2 applies directly, with $\varepsilon = h^p$, to the situation where a reversible numerical method of order p is used to discretize an integrable reversible system, or more generally, a reversible system with a KAM torus with diophantine frequencies ω. Here (4.1) corresponds to the time-h flow of the reversible system, and (4.3) represents the numerical map. This establishes the existence of invariant tori for reversible integrators, in perfect analogy to the symplectic counterpart Theorem X.6.2.

Concerning condition (X.6.3) we refer back to Sect. X.6.3, where it is shown that this condition is satisfied for a Cantor set of step sizes h if ω satisfies the diophantine condition (X.2.4).

XI.5 Exercises

1. This exercise shows that reversibility with respect to the particular involution $(u,v) \mapsto (u,-v)$ is not as special as it might seem at first glance.

(a) If the system $\dot{y} = f(y)$ is ρ-reversible (i.e., $f(\rho y) = -\rho f(y)$), then the transformed system $\dot{z} = T^{-1}f(Tz)$ is σ-reversible with $\sigma = T^{-1}\rho T$.
(b) Every linear involution ($\rho^2 = I$) is similar to a diagonal matrix with entries ± 1.

2. Consider the Toda lattice equations with an arbitrary number n of degrees of freedom and with periodic boundary conditions.
(a) Find all linear involutions ρ for which the system is ρ-reversible.
(b) Study for which ρ the eigenvalues of the matrix L are even functions of v.
(c) Investigate (numerically) the set of initial values for which all the assumptions of Theorem 1.3 are satisfied for some involution ρ.
Hint. Generalize the discussion for $n = 3$ in the Example 1.6.
3. A reversible system of the form

$$\begin{aligned} \dot{a} &= 0 \\ \dot{\theta} &= \omega(a, \theta) \end{aligned}$$

with ω an even function of $\theta \in \mathbb{T}^n$, also has a foliation of invariant tori. Consider reversible perturbations of such systems like in (2.1) and search for a reversibility-preserving transformation (2.3) that takes the perturbed system to the form

$$\begin{aligned} \dot{b} &= \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b, \varphi) + \varepsilon\mu(b, \varphi) + \mathcal{O}(\varepsilon^2) \end{aligned}$$

with μ even in φ. Write down the partial differential equations that the transformation must satisfy and discuss (sufficient) conditions for their solvability.
4. The torus $\{a = 0, \theta \in \mathbb{T}^n\}$ is invariant and carries a conditionally periodic flow with frequencies ω for reversible systems of the form $\dot{a} = \mathcal{O}(\|a\|)$, $\dot{\theta} = \omega + \mathcal{O}(\|a\|)$, which is more general than (2.13) in the differential equation for a. Discuss the difficulties that arise in trying to transform a reversible perturbation of such a system back to this form.
5. Apply an arbitrary (non-symmetric) Runge-Kutta method of even order $p = 2k$ to an integrable reversible system. Prove that under the assumptions of Theorem 3.1 the global error behaves for $t = nh$ like

$$y_n - y(t) = \mathcal{O}(th^p) + \mathcal{O}(t^2 h^{p+1}),$$

and the action variables like

$$I(y_n) - I(y_0) = \mathcal{O}(h^p) + \mathcal{O}(th^{p+1}).$$

Chapter XII. Dissipatively Perturbed Hamiltonian and Reversible Systems

Symplectic integrators also show a favourable long-time behaviour when they are applied to non-Hamiltonian perturbations of Hamiltonian systems. The same is true for symmetric methods applied to non-reversible perturbations of reversible systems. In this chapter we study the behaviour of numerical integrators when they are applied to dissipative perturbations of integrable systems, where only one invariant torus persists under the perturbation and becomes weakly attractive. The simplest example of such a system is Van der Pol's equation with small parameter, which has a single limit cycle in contrast to the infinitely many periodic orbits of the unperturbed harmonic oscillator.

XII.1 Numerical Experiments with Van der Pol's Equation

> One of the first such methods is the method of Van-der-Pol. [...] It should, however, be noted that in the formulation given by Van-der-Pol, approximation was effected by simple intuitive reasonings.
>
> (N.N. Bogoliubov & Y.A. Mitropolski 1961, p. 10f.)

Consider Van der Pol's equation

$$\begin{aligned} \dot{p} &= -q + \varepsilon(1-q^2)p \\ \dot{q} &= p \end{aligned} \tag{1.1}$$

with small positive ε, which is a perturbation of the harmonic oscillator. A symplectic change to polar coordinates $p = \sqrt{2a}\cos\theta,\ q = \sqrt{2a}\sin\theta$ puts the system into the form

$$\begin{aligned} \dot{a} &= \varepsilon\, 2a\cos^2\theta(1-2a\sin^2\theta) \\ \dot{\theta} &= 1 + \varepsilon\,\cos\theta\sin\theta(1-2a\sin^2\theta)\,. \end{aligned}$$

Since the angle θ evolves much faster than a, we may expect that the *averaged system*, which replaces the right-hand side functions by their angular averages, gives a good approximation:

$$\begin{aligned} \dot{a} &= \varepsilon\, a(1 - \tfrac{1}{2}a) \\ \dot{\theta} &= 1\,. \end{aligned}$$

Approximating by the averaged equation is the "method of Van-der-Pol" cited above, and the belief in the long-time validity of such an approximation is the *averaging principle*. The averaged differential equation for a has an unstable equilibrium at zero, and an asymptotically stable equilibrium at $a^* = 2$. The averaged system therefore has the circle $\{a^* = 2,\ \theta \in \mathbb{R} \bmod 2\pi\}$ as an attractive limit cycle. This suggests that the original Van der Pol equation has a nearby limit cycle, which is indeed the case.

Following the numerical experiment of Hairer & Lubich (1999), we solve the equation (1.1) with two initial values, $(p_0, q_0) = (0, 1.3)$ and $(p_0, q_0) = (0, 2.7)$, and with three numerical methods: the non-symplectic explicit and implicit Euler methods, and the symplectic Euler method. All of them have order 1. The numerical results are displayed in Fig. 1.1. For large step sizes (compared to the perturbation parameter ε), the non-symplectic methods give a completely wrong numerical solution, whereas that of the symplectic method is qualitatively correct. For smaller step sizes, the numerical solutions of the non-symplectic methods also show a limit cycle.

For the moment we explain these observations by "simple intuitive reasonings", that is, by the averaging principle and formal backward error analysis. The rigorous treatment is developed in the course of this chapter in a more general framework of perturbed integrable systems.

For a differential equation

$$\dot{y} = f(y) + \varepsilon g(y)\,,$$

the numerical solution y_n obtained by the explicit Euler method is the (formally) exact solution of a modified differential equation

$$\dot{\widetilde{y}} = f(\widetilde{y}) + \varepsilon g(\widetilde{y}) - \tfrac{1}{2} h f'(\widetilde{y}) f(\widetilde{y}) + \mathcal{O}(h^2 + \varepsilon h)\,.$$

For the Van der Pol equation in the above coordinates, the averaged modified equation becomes

$$\dot{\widetilde{a}} = h\widetilde{a} + \varepsilon \widetilde{a}(1 - \tfrac{1}{2}\widetilde{a}) + \ldots$$

which has approximately $\widetilde{a} = 2 + 2h/\varepsilon$ as an equilibrium. Hence, the limit cycle of the numerical solution of the explicit Euler method has approximate radius $2\sqrt{1 + h/\varepsilon}$ (Fig. 1.1) which is far from the correct value unless $h \ll \varepsilon$.

The implicit Euler discretization is adjoint to the explicit Euler method. Therefore, its modified differential equation is as above with h replaced by $-h$. In this case, the radius of the limit cycle is approximately $2\sqrt{1 - h/\varepsilon}$ (for $h < \varepsilon$), which again agrees very well with the pictures of Fig. 1.1.

For the symplectic Euler method, the modified differential equation for Van der Pol's equation is

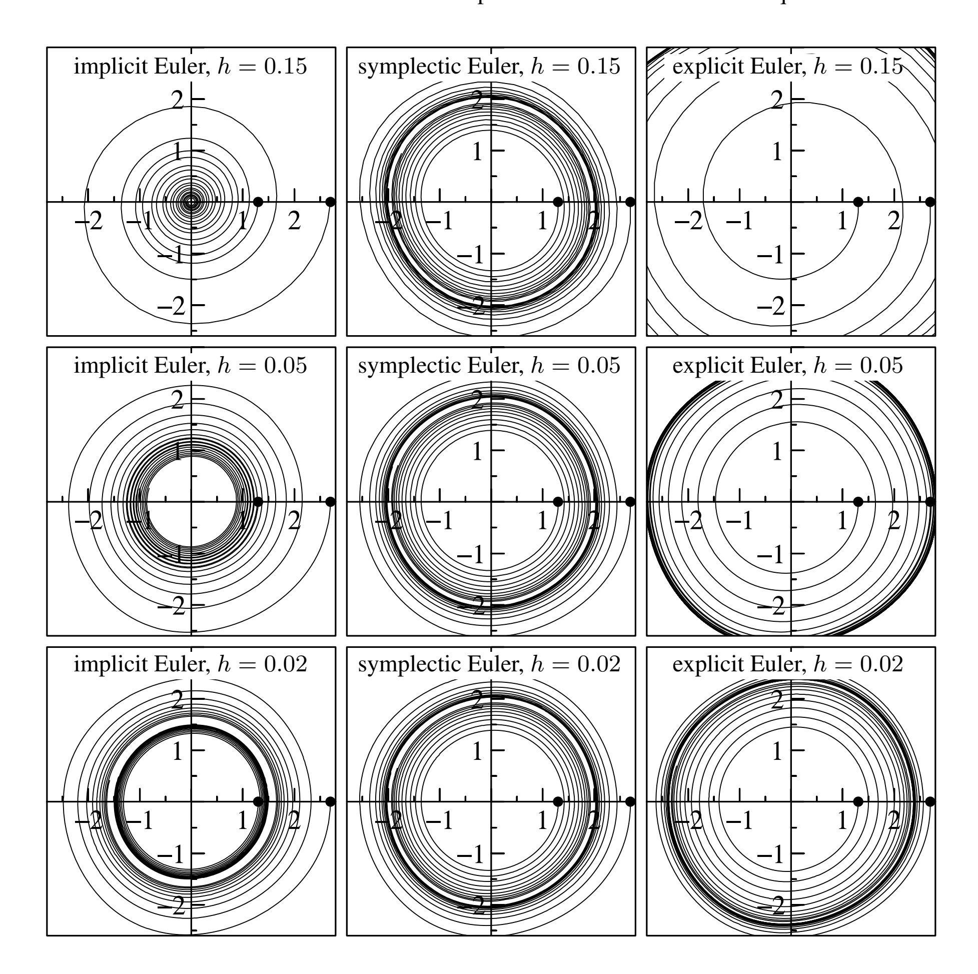


Fig. 1.1. Numerical experiments with Van der Pol's equation (1.1), $\varepsilon = 0.05$

$$\begin{aligned}\dot{\widetilde{p}} &= -\widetilde{q} + \varepsilon(1-\widetilde{q}^2)\widetilde{p} + \tfrac{1}{2}h\,\widetilde{p} + O(h^2+\varepsilon h)\\ \dot{\widetilde{q}} &= \widetilde{p} - \tfrac{1}{2}h\,\widetilde{q} + O(h^2+\varepsilon h).\end{aligned}$$

Here, the modified differential equation for the unperturbed harmonic oscillator is Hamiltonian (Theorem IX.3.1), and so all ε-independent terms in the averaged modified equation vanish:

$$\int_0^{2\pi} \frac{\partial H_j}{\partial \theta}(a,\theta)\,d\theta = 0.$$

Therefore, the radius of the limit cycle is of size $2 + \mathcal{O}(h)$ in accordance with Fig. 1.1.

XII.2 Averaging Transformations

> Le problème des oscillations non linéaires a actuellement une grande importance dans les domaines les plus divers de la technique et de la physique. Parmi les méthodes analytiques d'étude des oscillations non linéaires, la méthode asymptotique de développement en série par rapport à un paramètre petit est particulièrement efficace. Toute une série de monographies publiées en 1930–1938 par N. Krylov et N. Bogolioubov tant en russe qu'en français ont été consacrées à cette question, malheureusement ces ouvrages sont devenus aujourd'hui des raretés bibliographiques. Par ailleurs les méthodes exposées ont été largement développées depuis.
> (N. Bogolioubov & I. Mitropolski 1962, préface à la traduction française)

In this section we consider rather general perturbations of integrable systems. We study transformations that eliminate the dependence on the angles in the perturbation functions, up to arbitrary powers of the small perturbation parameter. The construction and properties of these "averaging " transformations are obtained by a slight extension of the arguments in Sections X.2 and XI.2.

XII.2.1 The Basic Scheme of Averaging

As in Sections X.2.1 and XI.2.1, we consider perturbations of an integrable system written in action-angle variables:

$$\begin{aligned} \dot{a} &= \varepsilon\, r(a,\theta) \\ \dot{\theta} &= \omega(a) + \varepsilon\, \rho(a,\theta) \end{aligned} \tag{2.1}$$

where ε is a small parameter and r, ρ are real-analytic in a neighbourhood of $\{a^*\} \times \mathbb{T}^d$. Unlike the situation of the previous chapters, we do not impose conditions that make the angular average

$$\overline{r}(a) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} r(a,\theta)\, d\theta \tag{2.2}$$

vanish identically. We look for a transformation to new variables (b, φ), of the form

$$\begin{aligned} a &= b + \varepsilon s(b,\varphi) \\ \theta &= \varphi + \varepsilon \sigma(b,\varphi)\ , \end{aligned} \tag{2.3}$$

which eliminates the dependence on the angles in the $\mathcal{O}(\varepsilon)$ terms of (2.1):

$$\begin{aligned} \dot{b} &= \varepsilon\, m(b) + \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon\, \mu(b) + \mathcal{O}(\varepsilon^2)\ . \end{aligned} \tag{2.4}$$

This is just a minor modification of the problem in Sect. XI.2.1. The equations that s and σ must satisfy, differ from (XI.2.5) and (XI.2.6) only in that the right-hand side $r(b, \varphi)$ of (XI.2.5) is replaced by $r(b, \varphi) - m(b)$, viz.,

$$\frac{\partial s}{\partial \varphi}(b,\varphi)\,\omega(b) \;=\; r(b,\varphi) - m(b) \tag{2.5}$$

$$\frac{\partial \sigma}{\partial \varphi}(b,\varphi)\,\omega(b) \;=\; \rho(b,\varphi) + \omega'(b)\,s(b,\varphi) - \mu(b)\;. \tag{2.6}$$

Necessary conditions for solvability are now

$$m(b) = \overline{r}(b)\;, \quad \mu(b) = \overline{\rho}(b)\;, \tag{2.7}$$

where the second equation corresponds to the choice $\overline{s}(b) = 0$. In other words, the leading terms in (2.4) are the angular averages of the perturbations in (2.1).

The equations (2.5), (2.6) are solvable for $b = b^*$ if $\omega(b^*)$ satisfies the diophantine condition (X.2.4). The “ultraviolet cutoff” argument of the proof of Lemma X.2.1 then shows that (2.4) holds uniformly as long as the solution remains in the ball $\|b - b^*\| \le c|\log\varepsilon|^{-\nu-1}$, with a sufficiently small constant c. This may hold over a very long time interval if the equation $\dot b = \varepsilon m(b)$ has a stable equilibrium in that ball.

XII.2.2 Perturbation Series

As in Sections X.2.2 and XI.2.2, the above construction extends to arbitrary finite order in ε. A transformation of the form (XI.2.9), which eliminates the angles in all terms up to order ε^{N-1}, is sought for:

$$\begin{aligned} \dot b \;&=\; \varepsilon m_1(b) + \varepsilon^2 m_2(b) + \ldots + \varepsilon^{N-1} m_{N-1}(b) + \varepsilon^N r_N(b,\varphi) \\ \dot\varphi \;&=\; \omega(b) + \varepsilon\mu_1(b) + \varepsilon^2\mu_2(b) + \ldots + \varepsilon^{N-1}\mu_{N-1}(b) + \varepsilon^N \rho_N(b,\varphi)\;. \end{aligned} \tag{2.8}$$

The equations determining the transformation are a slight modification of (XI.2.11) and (XI.2.12): on the right-hand side of (XI.2.11), $p_j(b,\varphi)$ is replaced by the difference $p_j(b,\varphi) - m_j(b)$, with $m_j(b) = \overline{p}_j(b)$. We then have the following variant of Lemmas X.2.1 and XI.2.1.

Lemma 2.1. *Let the right-hand side functions of (2.1) be real-analytic in a neighbourhood of $\{b^*\} \times \mathbb{T}^d$. Suppose that $\omega(b^*)$ satisfies the diophantine condition (X.2.4) with exponent ν. For any fixed $N \ge 2$, there are positive constants ε_0, c, C such that the following holds for $|\varepsilon| \le \varepsilon_0$: there exists a real-analytic change of coordinates $(a,\theta) \mapsto (b,\varphi)$ which transforms (2.1) to (2.8) with*

$$\begin{aligned} &\|m_j(b)\| \le C/\delta^{j-1}\;, && \|\mu_j(b)\| \le C/\delta^{j-1} \\ &\|r_N(b,\varphi)\| \le C/\delta^{N-1}\;, && \|\rho_N(b,\varphi)\| \le C/\delta^{N-1} \end{aligned} \qquad \textit{for} \quad \|b - b^*\| \le \delta\;,$$

where

$$\delta = c|\log\varepsilon|^{-\nu-1}\;. \tag{2.9}$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$-close to the identity: $\|(a,\theta) - (b,\varphi)\| \le C\varepsilon$ holds for (a,θ) and (b,φ) related by the above coordinate transform, for $\|b - b^*\| \le \delta$ and for φ in an ε-independent complex neighbourhood of $\mathbb{T}^d$.

The constants ε_0, c, C depend on N, d, γ, ν and on bounds of ω, r, ρ on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$.

Proof. The proof uses again the ultraviolet cutoff argument of the proof of Lemma X.2.1. This makes all the functions $s_i, \sigma_i, m_i, \mu_i$ real-analytic in b for $\|b-b^*\| \leq 2\delta$ and of φ in an ε-independent complex neighbourhood of $\mathbb{T}^d$. The powers of δ in the denominators of the estimates come from the presence of terms $\partial s_j/\partial b$, $\partial\sigma_j/\partial b$ in $p_i(b,\varphi)$ and $\pi_i(b,\varphi)$ of (XI.2.11) and (XI.2.12) and from Cauchy's estimates applied to s_j, σ_j on $\|b-b^*\| \leq 2\delta$. □

XII.3 Attractive Invariant Manifolds

> Theorems on invariant manifolds for maps have been proved many times for many different settings. The first results were obtained by Hadamard (1901) and Perron (1929). [...] Our aim was to derive a global invariant manifold result with conditions that are easy to verify for the applications in mind. (K. Nipp & D. Stoffer 1992)

In this section we give results on the existence and properties of attractive invariant manifolds of maps, with a very explicit handling of constants. These results are due to Kirchgraber, Lasagni, Nipp & Stoffer (1991) and Nipp & Stoffer (1992). They will allow us to understand the weakly attractive closed curves that we observed in Sect. XII.1. Beyond that particular example, these results are extremely useful for studying the long-time behaviour of numerical discretizations in a great variety of applications; see Nipp & Stoffer (1995, 1996) and Lubich (2001) and references therein, and also Stuart & Humphries (1996) for a related invariant manifold theorem and its use in analyzing the dynamics of numerical integrators for non-conservative problems.

Consider a map $\Phi : X \times Y \to X \times Y$ defined on the Cartesian product of a Banach space X and a closed bounded subset Y of another Banach space. We write $\Phi(x,y) = (\widehat{x},\widehat{y})$ with

$$\begin{aligned} \widehat{x} &= x + f(x,y) \\ \widehat{y} &= \phantom{x+{}} g(x,y) \, . \end{aligned} \tag{3.1}$$

We assume that f and g are Lipschitz bounded, with Lipschitz constants L_{xx}, L_{xy} and L_{yx}, L_{yy} with respect to x, y. If these Lipschitz constants are sufficiently small, then the map Φ has an attractive invariant manifold. More precisely, there is the following result, stated without proof by Kirchgraber, Lasagni, Nipp & Stoffer (1991) and proved in a more general setting by Nipp & Stoffer (1992).

Theorem 3.1. *In the above situation, if*

$$L_{xx} + L_{yy} + 2\sqrt{L_{xy}L_{yx}} < 1 \, , \tag{3.2}$$

then there exists a function $s : X \to Y$, which is Lipschitz bounded with the constant $\lambda = 2L_{yx}/(1-L_{xx}-L_{yy})$, such that

$$\mathcal{M} = \{(x,s(x)) : x \in X\} \ \textit{ is invariant under } \Phi.$$

$\mathcal{M}$ attracts orbits of Φ with the attractivity factor $\rho = \lambda L_{xy} + L_{yy} < 1$, that is, $\|\widehat{y} - s(\widehat{x})\| \le \rho \, \|y - s(x)\|$ holds for all $(x, y) \in X \times Y$.

Proof. (a) We search for a function $s : X \to Y$ such that for $(\widehat{x}, \widehat{y}) = \Phi(x, y)$, the relation $y = s(x)$ implies also $\widehat{y} = s(\widehat{x})$. For an arbitrary function $\sigma : X \to Y$, we first study which relation holds between $\widehat{x}$ and $\widehat{y}$ if $y = \sigma(x)$. To write $\widehat{y}$ as a function of $\widehat{x}$, we need a bijective correspondence between x and $\widehat{x}$ via the first equation of (3.1). By the Banach fixed-point theorem, the equation

$$\widehat{x} = x + f(x, \sigma(x)) \;\text{ has a unique solution }\; x = u_\sigma(\widehat{x})$$

for every $\widehat{x} \in X$ if $x \mapsto f(x, \sigma(x))$ is a contraction. This is the case if σ has the Lipschitz constant λ and

$$L_{xx} + L_{xy}\lambda < 1 \,. \tag{3.3}$$

We then obtain $\widehat{y} = \widehat{\sigma}(\widehat{x})$ from the following scheme:

$$\begin{array}{ccc} x = u_\sigma(\widehat{x}) & \longleftarrow & \widehat{x} \\ \Big\downarrow \sigma & & \\ y = \sigma(x) & \longrightarrow & \widehat{y} = g(x, y) \end{array}$$

That is, we set $\widehat{y} = \widehat{\sigma}(\widehat{x}) = g(u_\sigma(\widehat{x}), \sigma(u_\sigma(\widehat{x})))$. By construction, $(\widehat{x}, \widehat{y}) = \Phi(x, y)$. Under condition (3.3), the function $u_\sigma : X \to X$ is Lipschitz bounded by $\mu = 1/(1 - L_{xx} - L_{xy}\lambda)$. Consequently, the function $\widehat{\sigma} : X \to Y$ is Lipschitz bounded by $(L_{yx} + L_{yy}\lambda)\mu$. The condition that the transformed function $\widehat{\sigma}$ is again Lipschitz bounded by the same λ as σ, therefore reads

$$\frac{L_{yx} + L_{yy}\lambda}{1 - L_{xx} - L_{xy}\lambda} \le \lambda \,, \tag{3.4}$$

or equivalently,

$$L_{xy}\lambda^2 - (1 - L_{xx} - L_{yy})\lambda + L_{yx} \le 0 \,.$$

Under condition (3.2), there exists a non-empty real interval of values λ satisfying this quadratic inequality. In particular, (3.4) then holds for

$$\lambda = \frac{2L_{yx}}{1 - L_{xx} - L_{yy}} \,. \tag{3.5}$$

(This is close to the smallest possible value of λ if $2\sqrt{L_{xy}L_{yx}} \ll 1 - L_{xx} - L_{yy}$.) It is easily checked that (3.2) and (3.5) imply (3.3).

Under conditions (3.3) and (3.4), the transformation $H : \sigma \mapsto \widehat{\sigma}$, which is called a *Hadamard graph transform*, maps the set of functions

$$S = \{\sigma : X \to Y \mid \; \sigma \text{ is Lipschitz bounded by } \lambda\}$$

into itself, i.e.,

$$H : S \to S : \sigma \mapsto \widehat{\sigma} .$$

S is a closed subset of $C(X, Y)$, the Banach space of continuous functions from X to the bounded closed set Y, equipped with the supremum norm $\|\sigma\|_\infty = \sup_{x \in X} \|\sigma(x)\|$. If H is a contraction, then the Banach fixed-point theorem tells us that there is a unique function $s \in S$ with $\widehat{s} = s$. By construction, this means that if $(\widehat{x}, \widehat{y}) = \Phi(x, y)$ and $y = s(x)$, then also $\widehat{y} = s(\widehat{x})$. The graph $\mathcal{M} = \{(x, s(x)) : x \in X\}$ is then an invariant manifold for the map Φ.

(b) We now show that H is already a contraction under condition (3.2). Let σ_0, σ_1 be two arbitrary functions in S, and $\widehat{x} \in X$. With $x_i = u_{\sigma_i}(\widehat{x})$,

$$\begin{aligned}
&\|H\sigma_1(\widehat{x}) - H\sigma_0(\widehat{x})\| = \|g(x_1, \sigma_1(x_1)) - g(x_0, \sigma_0(x_0))\| \\
&\le \|g(x_1, \sigma_1(x_1)) - g(x_1, \sigma_0(x_1))\| + \|g(x_1, \sigma_0(x_1)) - g(x_0, \sigma_0(x_0))\| \\
&\le L_{yy} \|\sigma_1 - \sigma_0\|_\infty + (L_{yx} + L_{yy}\lambda) \|x_1 - x_0\| .
\end{aligned}$$

By definition, $\widehat{x} = x_i + f(x_i, \sigma_i(x_i))$ for $i = 0, 1$. Subtracting these two equations yields similarly

$$\begin{aligned}
&\|x_1 - x_0\| \le \|f(x_1, \sigma_1(x_1)) - f(x_0, \sigma_0(x_0))\| \\
&\le \|f(x_1, \sigma_1(x_1)) - f(x_1, \sigma_0(x_1))\| + \|f(x_1, \sigma_0(x_1)) - f(x_0, \sigma_0(x_0))\| \\
&\le L_{xy} \|\sigma_1 - \sigma_0\|_\infty + (L_{xx} + L_{xy}\lambda) \|x_1 - x_0\| .
\end{aligned}$$

Hence,

$$\|x_1 - x_0\| \le \frac{L_{xy}}{1 - L_{xx} - L_{xy}\lambda} \|\sigma_1 - \sigma_0\|_\infty .$$

Combining both inequalities and recalling (3.4), we obtain

$$\|H\sigma_1 - H\sigma_0\|_\infty \le (L_{yy} + \lambda L_{xy}) \|\sigma_1 - \sigma_0\|_\infty .$$

Since the inequality

$$L_{yy} + \lambda L_{xy} < 1 \tag{3.6}$$

is satisfied by the λ of (3.5) under condition (3.2), H is indeed a contraction.

(c) It remains to show that the invariant manifold $\mathcal{M}$ is attractive. With $(\widehat{x}, \widehat{y}) = \Phi(x, y)$, we write

$$\begin{aligned}
&\widehat{y} - s(\widehat{x}) = g(x, y) - s(x + f(x, y)) \\
&= \Big(g(x, y) - g(x, s(x))\Big) + \Big(s(x + f(x, s(x))) - s(x + f(x, y))\Big) .
\end{aligned}$$

Here we used the identity

$$s(x + f(x, s(x))) = \widehat{s}(x + f(x, s(x))) = g(x, s(x)) ,$$

which holds because $\widehat{s} = s$ and by construction of the Hadamard transform. It follows that

$$\|\widehat{y} - s(\widehat{x})\| \le (L_{yy} + \lambda L_{xy}) \|y - s(x)\| ,$$

which together with (3.6) yields the result. □

Next we study the effect of a perturbation of the map on the invariant manifold.

Theorem 3.2. *Consider maps Φ_0, $\Phi_1 : X \times Y \to X \times Y$ both of which satisfy the conditions of Theorem 3.1 with the same Lipschitz constants $L_{xx}, L_{xy}, L_{yx}, L_{yy}$. Let s_0 and s_1 be the functions defining the attractive invariant manifolds $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. If the bound*

$$\|\Phi_1(x,y) - \Phi_0(x,y)\| \le \delta \quad \textit{for} \quad (x,y) \in \mathcal{M}_0$$

holds in the norm $\|(x,y)\| = \lambda \|x\| + \|y\|$ on $X \times Y$, then

$$\|s_1(x) - s_0(x)\| \le \frac{\delta}{1-\rho} \quad \textit{for} \quad x \in X\,.$$

(Here λ and ρ are defined as in Theorem 3.1.)

Proof. The proof is similar to part (b) of the previous proof. Let $\widehat{x} \in X$. For $i = 0,1$, we have $s_i(\widehat{x}) = g_i(x_i, s_i(x_i))$ with x_i defined by the equation $\widehat{x} = x_i + f_i(x_i, s_i(x_i))$. We estimate

$$\begin{aligned}
\|s_1(\widehat{x}) - s_0(\widehat{x})\| &\le \|g_1(x_1, s_1(x_1)) - g_1(x_1, s_0(x_1))\| \\
&+ \|g_1(x_1, s_0(x_1)) - g_1(x_0, s_0(x_0))\| \\
&+ \|g_1(x_0, s_0(x_0)) - g_0(x_0, s_0(x_0))\| \\
&\le L_{yy}\|s_1 - s_0\|_\infty + (L_{yx} + L_{yy}\lambda)\,\|x_1 - x_0\| \\
&+ \|g_1(x_0, s_0(x_0)) - g_0(x_0, s_0(x_0))\|
\end{aligned}$$

and in the same way

$$\begin{aligned}
\|x_1 - x_0\| &\le \|f_1(x_1, s_1(x_1)) - f_1(x_1, s_0(x_1))\| \\
&+ \|f_1(x_1, s_0(x_1)) - f_1(x_0, s_0(x_0))\| \\
&+ \|f_1(x_0, s_0(x_0)) - f_0(x_0, s_0(x_0))\| \\
&\le L_{xy}\|s_1 - s_0\|_\infty + (L_{xx} + L_{xy}\lambda)\,\|x_1 - x_0\| \\
&+ \|f_1(x_0, s_0(x_0)) - f_0(x_0, s_0(x_0))\|\,.
\end{aligned}$$

Inserting the second bound into the first one and using (3.4) and the assumed bound on $\Phi_1 - \Phi_0$ gives

$$\|s_1 - s_0\|_\infty \le (L_{yy} + \lambda L_{xy})\,\|s_1 - s_0\|_\infty + \delta\,,$$

which implies the result. □

XII.4 Weakly Attractive Invariant Tori of Perturbed Integrable Systems

We assume that the perturbation is dissipative such that one torus persists under the perturbation and gets attractive.

Our analysis is done by the method of averaging. The problem of this section is classical, see e.g. Bogoliubov & Mitropolski (1961), Kirchgraber & Stiefel (1978). (D. Stoffer 1998)

In the example of the Van der Pol equation, we have seen that only one of the periodic orbits of the harmonic oscillator persists under the small nonlinear perturbation and becomes an attractive limit cycle. More generally, we consider perturbations of integrable systems

$$\begin{aligned} \dot a &= \varepsilon\, r(a,\theta) \\ \dot\theta &= \omega(a) + \varepsilon\, \rho(a,\theta) \end{aligned} \tag{4.1}$$

where (locally) just one invariant torus survives the perturbation and attracts nearby solutions. Using the results of the two previous sections, it will be shown that this situation occurs if, at some point a^* where the frequencies $\omega_i(a^*)$ are diophantine, the angular average $\overline{r}(a^*)$ is small and its Jacobian matrix

$$A = \overline{r}\,'(a^*)$$

has all eigenvalues with negative real part.

The following theorem is a slight modification of a result of Stoffer (1998). Early versions of it are much older; see the citations above. The origins of the problem can be traced back to the work of Van der Pol (1927) and Krylov & Bogoliubov (1934).

Here we assume the following: $\omega(a^*)$ satisfies the diophantine condition (X.2.4) with exponent ν. The perturbation functions $r(a,\theta)$ and $\rho(a,\theta)$ are real-analytic on a fixed complex neighbourhood of $\{a^*\}\times\mathbb{T}^d$ and bounded independently of ε (though they may depend on ε). In some norm $\|\cdot\|$ on $\mathbb{R}^d$ and its induced matrix norm, the bounds

$$\|\overline{r}(a^*)\| \;\le\; C|\log\varepsilon|^{-2(\nu+1)} \tag{4.2}$$

$$\|e^{tA}\| \;\le\; e^{-t\alpha} \qquad \text{for} \quad t>0 \tag{4.3}$$

hold with some constants C and $\alpha>0$.

Theorem 4.1. *Under the above conditions, for sufficiently small $\varepsilon>0$, the system (4.1) has an invariant torus $\mathcal{T}_\varepsilon$ which attracts an $\mathcal{O}(|\log\varepsilon|^{-\nu-1})$-neighbourhood of $\{a^*\}\times\mathbb{T}^d$ with an exponential rate proportional to ε.*

Proof. The proof combines Lemma 2.1 and Theorem 3.1. For convenience we assume $a^*=0$ in the following. Lemma 2.1 (with $N=3$) gives us a change of coordinates $(a,\theta)\mapsto(b,\varphi)$, $\mathcal{O}(\varepsilon)$-close to the identity, such that for $\|b\|\le\delta$ with $\delta=c|\log\varepsilon|^{-\nu-1}$ of (2.9),

$$\begin{aligned} \dot{b} &= \varepsilon m_1(b) + \varepsilon^2 m_2(b) + \mathcal{O}(\varepsilon^3/\delta^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon\mu_1(b) + \varepsilon^2\mu_2(b) + \mathcal{O}(\varepsilon^3/\delta^2)\ . \end{aligned} \tag{4.4}$$

Since $m_1(b) = \overline{r}(b) = Ab + \mathcal{O}(\delta^2)$ by (4.2), this system is of the form

$$\begin{aligned} \dot{b} &= \varepsilon Ab + \mathcal{O}(\varepsilon\delta^2) \\ \dot{\varphi} &= \omega(b) + \mathcal{O}(\varepsilon)\ . \end{aligned}$$

Similarly, the corresponding variational equation is of the form

$$\begin{pmatrix} \dot{B} \\ \dot{\Phi} \end{pmatrix} = \begin{pmatrix} \varepsilon A + \mathcal{O}(\varepsilon\delta) & \mathcal{O}(\varepsilon^3/\delta^2) \\ \mathcal{O}(1) & \mathcal{O}(\varepsilon^3/\delta^2) \end{pmatrix} \begin{pmatrix} B \\ \Phi \end{pmatrix}\ .$$

These relations and condition (4.3) imply that, for sufficiently small ε and for any fixed $\tau > 0$, the time-τ flow of (4.1) maps the strip $D = \{(b,\varphi) : \|b\| \le \frac{1}{2}\delta, \varphi \in \mathbb{T}^d\}$ into itself, and the following bounds hold for the derivatives of the solution with respect to the initial values:

$$\begin{aligned} &\left\| \frac{\partial b(\tau)}{\partial b(0)} \right\| \le L_{bb} = e^{-\tau\varepsilon\alpha} + \mathcal{O}(\varepsilon\delta)\ , && \left\| \frac{\partial b(\tau)}{\partial \varphi(0)} \right\| \le L_{b\varphi} = \mathcal{O}(\varepsilon^3/\delta^2) \\ &\left\| \frac{\partial \varphi(\tau)}{\partial b(0)} \right\| \le L_{\varphi b} = \mathcal{O}(1)\ , && \left\| \frac{\partial \varphi(\tau)}{\partial \varphi(0)} - I \right\| \le L_{\varphi\varphi} = \mathcal{O}(\varepsilon^3/\delta^2)\ . \end{aligned} \tag{4.5}$$

Hence, for sufficiently small ε,

$$L_{\varphi\varphi} + L_{bb} + 2\sqrt{L_{\varphi b}L_{b\varphi}} \le e^{-\tau\varepsilon\alpha/2} < 1\ .$$

Theorem 3.1 (and Exercise 1) used with φ, b in the roles of x, y now shows that the time-τ flow has an attractive invariant torus $\{(s(\varphi),\varphi) : \varphi \in \mathbb{T}^d\}$, where $s : \mathbb{T}^d \to \{\|b\| \le \frac{1}{2}\delta\}$ is Lipschitz bounded by $\lambda = 2L_{b\varphi}/(1 - L_{\varphi\varphi} - L_{bb}) = \mathcal{O}(\varepsilon^3/\delta^2)$. This invariant torus attracts orbits of the time-τ flow map in the strip D with the attractivity factor $\lambda L_{\varphi b} + L_{bb} \le e^{-\tau\varepsilon\alpha/2}$. As Exercise 2 shows, the torus is actually invariant for the differential equation (4.1). □

XII.5 Weakly Attractive Invariant Tori of Numerical Integrators

Does the attractive invariant torus of Theorem 4.1 persist under numerical discretization of the perturbed integrable system? This question was first studied by Stoffer (1998) who worked directly with the discrete equations in his analysis. Here we take up the approach of Hairer & Lubich (1999) where the problem was studied by combining backward error analysis and perturbation theory, similar to what was done in the two preceding chapters.

XII.5.1 Modified Equations of Perturbed Differential Equations

Below we need to use backward error analysis for the numerical solution of a perturbed differential equation

$$\dot{y} = f(y) + \varepsilon g(y, \varepsilon) , \qquad y(0) = y_0 \tag{5.1}$$

with real-analytic functions f and g and small parameter ε. We consider applying a one-step method $y_1 = \Phi_h^\varepsilon(y_0)$ of order $p \geq 1$ with step size $h > 0$. The associated modified differential equations constructed in Chap. IX are then of the form

$$\dot{\widetilde{y}} = \widetilde{f}(\widetilde{y}) + \varepsilon \widetilde{g}(\widetilde{y}, \varepsilon) , \qquad \widetilde{y}(0) = y_0 \tag{5.2}$$

with suitably truncated series

$$\begin{aligned} \widetilde{f}(y) &= f(y) + h^p f_{p+1}(y) + \ldots + h^{N-1} f_N(y) \\ \widetilde{g}(y, \varepsilon) &= g(y, \varepsilon) + h^p g_{p+1}(y, \varepsilon) + \ldots + h^{N-1} g_N(y, \varepsilon) , \end{aligned} \tag{5.3}$$

where the functions f_j are independent of ε, h, N, whereas the functions g_j are allowed to depend on ε. The following adapts Theorem IX.7.6 to the above situation.

Theorem 5.1. *Let $f(y) + \varepsilon g(y, \varepsilon)$ be real-analytic (in y and ε) and bounded by M for $y \in B_{2R}(y_0)$ and for all complex ε with $|\varepsilon| \leq \varepsilon_0$. Let the coefficients of the Taylor series (in h) of the numerical method be analytic in $B_R(y_0)$ with bounds (IX.7.5) for $|\varepsilon| \leq \varepsilon_0$. Then, there exists $h_0 > 0$ (proportional to R/M), such that for $h \leq h_0/4$ and for $N = N(h)$ the largest integer with $hN \leq h_0$, the difference between the numerical solution $y_1 = \Phi_h^\varepsilon(y_0)$ and the exact solution $\widetilde{\varphi}_{N,t}^\varepsilon(y_0)$ of the truncated modified equation (5.2)-(5.3) satisfies*

$$\|\Phi_h^\varepsilon(y_0) - \widetilde{\varphi}_{N,h}^\varepsilon(y_0)\| \leq C h\, e^{-h_0/h}.$$

The functions $\widetilde{f}$ and $\widetilde{g}$ of (5.3) are real-analytic in $B_R(y_0)$ with

$$\|\widetilde{f}(y) - f(y)\| \leq Ch^p , \quad \|\widetilde{g}(y, \varepsilon) - g(y, \varepsilon)\| \leq Ch^p$$

for $y \in B_{R/2}(y_0)$ and $|\varepsilon| \leq \varepsilon_0$. The constants C are independent of $h \leq h_0/4$ and $|\varepsilon| \leq \varepsilon_0$.

Proof. The exponentially small estimate for $\Phi_h^\varepsilon(y_0) - \widetilde{\varphi}_{N,h}^\varepsilon(y_0)$ is that of Theorem IX.7.6 applied to the differential equation (5.1). The $\mathcal{O}(h^p)$ bound for $\widetilde{f}(y) - f(y)$ is the estimate (IX.7.14) applied to $\dot{y} = f(y)$. By applying that estimate to (5.1), a bound of the same type is obtained for $(\widetilde{f}(y) + \varepsilon \widetilde{g}(y, \varepsilon)) - (f(y) + \varepsilon g(y))$, uniformly for all complex ε in the complex disk $|\varepsilon| \leq \varepsilon_0$. For any fixed $y \in B_{R/2}(y_0)$, the difference

$$\widetilde{g}(y, \varepsilon) - g(y, \varepsilon) = \frac{1}{\varepsilon} \Big([(\widetilde{f}(y) + \varepsilon \widetilde{g}(y, \varepsilon)) - (f(y) + \varepsilon g(y, \varepsilon))] - [\widetilde{f}(y) - f(y)] \Big)$$

is an analytic function of ε in the complex disk $|\varepsilon| \leq \varepsilon_0$, which is bounded by $\mathcal{O}(h^p)$ for $|\varepsilon| = \varepsilon_0$. By the maximum principle, the same bound then holds for $|\varepsilon| \leq \varepsilon_0$. □

XII.5.2 Symplectic Methods

We apply a symplectic integrator with step size h to a real-analytic perturbed integrable Hamiltonian system in coordinates (p, q),

$$\begin{aligned} \dot{p} &= -\frac{\partial H}{\partial q}(p,q) + \varepsilon k(p,q) \\ \dot{q} &= \frac{\partial H}{\partial p}(p,q) + \varepsilon \ell(p,q) \, . \end{aligned} \tag{5.4}$$

We assume that the unperturbed system ($\varepsilon = 0$) is a completely integrable system which satisfies the conditions of the Arnold–Liouville theorem, Theorem X.1.6. Hence, there exists a transformation to action-angle variables for the

$$\textit{integrable system:} \;\; (p,q) \mapsto (a,\theta) \;\; \text{by Theorem X.1.6.}$$

This change of coordinates transforms the integrable system to the equations $\dot{a} = 0$, $\dot{\theta} = \omega(a)$, and it transforms (5.4) to a system (4.1), for which we assume (4.2), (4.3) and the diophantine condition (X.2.4) with exponent ν for $\omega(a^*)$. The following theorem is a variant of results in Stoffer (1998) and Hairer & Lubich (1999). It shows that for symplectic methods, the invariant torus persists under a very mild restriction on the step size. For non-symplectic methods, this would require step sizes h with $h^p \ll \varepsilon$ (see Exercise 5).

Theorem 5.2. *Let a symplectic numerical integrator of order p be applied to a perturbed integrable Hamiltonian system (5.4) which satisfies the conditions stated above. Then, there exist $\varepsilon_0 > 0$ and $c_0 > 0$ such that, for $0 < \varepsilon \le \varepsilon_0$ and for step sizes $h > 0$ satisfying*

$$h^p \le c_0 \, |\log \varepsilon|^{-\kappa} \tag{5.5}$$

with $\kappa = \max(\nu + d + 1, p)$, the numerical method has an attractive invariant torus $\mathcal{T}_{\varepsilon,h}$. This torus is $\mathcal{O}(h^p)$ close to the invariant torus $\mathcal{T}_\varepsilon$ of (5.4). It attracts an $\mathcal{O}(|\log \varepsilon|^{-2\kappa})$ neighbourhood with an exponential rate proportional to ε, uniformly in h.

Remark 5.3. The exponent $\nu+d+1$ comes from Lemma X.4.1. It could be reduced to $\nu + 1$ by using Rüssmann's estimates in place of that lemma; cf. the remark after Lemma X.4.1.

Proof of Theorem 5.2. The proof combines backward error analysis (Theorem IX.3.1 and Theorem 5.1), perturbation theory (Theorem X.4.4 and Lemma 2.1), and the invariant manifold theorem (Theorem 3.1).

(a) We begin by considering the symplectic method applied to the integrable Hamiltonian system (5.4) with $\varepsilon = 0$. This leads us back to the questions of Chap. X. We use backward error analysis and recall (Theorem IX.3.1) that the modified equation is again Hamiltonian and an $\mathcal{O}(h^p)$ perturbation of the integrable system, both in the (p, q) and the (a, θ) variables. We transform variables for the

modified equation of the integrable system: $(a,\theta) \mapsto (\widetilde{a},\widetilde{\theta})$ by Theorem X.4.4,

with h^p in the role of the perturbation parameter. By (X.4.1) with N proportional to $|\log\varepsilon|$, and by condition (5.5) with a sufficiently small c_0, the modified equations in these variables become

$$\begin{aligned} \dot{\widetilde{a}} &= \mathcal{O}(\varepsilon^3) \\ \dot{\widetilde{\theta}} &= \widetilde{\omega}(\widetilde{a}) + \mathcal{O}(\varepsilon^3) \end{aligned} \qquad \text{for} \quad \|\widetilde{a}-a^*\| \le c^*|\log\varepsilon|^{-2\kappa} ,$$

with $\widetilde{\omega}(\widetilde{a}) = \omega(\widetilde{a}) + \mathcal{O}(h^p)$. Moreover, the transformation $(a,\theta) \mapsto (\widetilde{a},\widetilde{\theta})$ is $\mathcal{O}(h^p)$ close to the identity.

(b) The modified equations of the perturbed system, written in the $(\widetilde{a},\widetilde{\theta})$ variables, become

$$\begin{aligned} \dot{\widetilde{a}} &= \varepsilon\widetilde{r}(\widetilde{a},\widetilde{\theta}) + \mathcal{O}(\varepsilon^3) \\ \dot{\widetilde{\theta}} &= \widetilde{\omega}(\widetilde{a}) + \varepsilon\widetilde{\rho}(\widetilde{a},\widetilde{\theta}) + \mathcal{O}(\varepsilon^3) \end{aligned} \qquad \text{for} \quad \|\widetilde{a}-a^*\| \le c^*|\log\varepsilon|^{-2\kappa} , \tag{5.6}$$

where $\widetilde{r}(\widetilde{a},\widetilde{\theta}) = r(\widetilde{a},\widetilde{\theta}) + \mathcal{O}(h^p)$ and $\widetilde{\rho}(\widetilde{a},\widetilde{\theta}) = \rho(\widetilde{a},\widetilde{\theta}) + \mathcal{O}(h^p)$ by Theorem 5.1. Consider now these equations with the $\mathcal{O}(\varepsilon^3)$ terms dropped. We change variables for the

modified equation of the perturbed system: $(\widetilde{a},\widetilde{\theta}) \mapsto (\widetilde{b},\widetilde{\varphi})$ by Lemma 2.1.

(Note Exercise 4 with $\widetilde{\omega}(a^*) = \omega(a^*) + \mathcal{O}(h^p)$ and (5.5).) The system (5.6) is transformed to the form of (4.4),

$$\begin{aligned} \dot{\widetilde{b}} &= \varepsilon\widetilde{m}(\widetilde{b}) + \mathcal{O}(\varepsilon^3/\delta^2) \\ \dot{\widetilde{\varphi}} &= \widetilde{\omega}(\widetilde{b}) + \varepsilon\widetilde{\mu}(\widetilde{b}) + \mathcal{O}(\varepsilon^3/\delta^2) \end{aligned} \tag{5.7}$$

with $\delta = c^*|\log\varepsilon|^{-2\kappa}$, and where $\widetilde{m}(\widetilde{b}) = \overline{\widetilde{r}}(\widetilde{b}) + \mathcal{O}(\varepsilon/\delta) = \overline{r}(\widetilde{b}) + \mathcal{O}(h^p) + \mathcal{O}(\varepsilon/\delta)$, and also the Jacobian of $\widetilde{m}$ at a^* is close to that of $\overline{r}$, so that it satisfies again (4.3), at least with α replaced by $\alpha/2$. In the same way as in the proof of Theorem 4.1 and with the same Lipschitz constants as in (4.5), we now obtain an attractive invariant torus of the modified equation of the perturbed system. The time-h flow of this equation is an exponentially small (in $1/h$) Lipschitz perturbation of the numerical one-step map, so that under condition (5.5) it is an $\mathcal{O}(\varepsilon^3)$ perturbation. Therefore, Theorem 3.1 yields an invariant torus $\mathcal{T}_{\varepsilon,h}$ of the numerical method.

(c) It remains to bound the distance between the tori $\mathcal{T}_{\varepsilon,h}$ and $\mathcal{T}_\varepsilon$. We recall that $\mathcal{T}_\varepsilon$ was obtained by a transformation of the

perturbed system: $(a,\theta) \mapsto (b,\varphi)$ by Lemma 2.1,

which puts (4.1) into the form (4.4). We thus have the transformations

$$\begin{array}{ccc} (a,\theta) & \overset{\varepsilon}{\longrightarrow} & (b,\varphi) \\ h^p \Big\downarrow & & \\ (\widetilde{a},\widetilde{\theta}) & \overset{\varepsilon}{\longrightarrow} & (\widetilde{b},\widetilde{\varphi}) \end{array}$$

where the symbols h^p and ε indicate that the transformation is $\mathcal{O}(h^p)$ or $\mathcal{O}(\varepsilon)$ close to the identity. By the construction of Lemma 2.1, the composed transformation $(b,\varphi) \mapsto (\widetilde{b},\widetilde{\varphi})$ is $\mathcal{O}(h^p)$ close to the identity and moreover, the right-hand sides of (4.4) and (5.7) differ by $\mathcal{O}(\varepsilon h^p)$. Theorem 3.2 (with $\rho = e^{-\varepsilon\tau\alpha/2}$) now shows that the functions $s_{\varepsilon,h}$ and s_ε defining $\mathcal{T}_{\varepsilon,h}$ and $\mathcal{T}_\varepsilon$, respectively, differ by $\mathcal{O}(h^p)$. This yields the desired distance bound. □

XII.5.3 Symmetric Methods

A result analogous to the theorem of the previous subsection holds for reversible methods applied to perturbed reversible systems

$$\begin{aligned} \dot{u} &= f(u,v) + \varepsilon k(u,v) \\ \dot{v} &= g(u,v) + \varepsilon \ell(u,v) \end{aligned}$$

where the unperturbed system ($\varepsilon = 0$) is a real-analytic integrable reversible system. If the perturbed system, written in action-angle variables of the unperturbed system, satisfies the conditions of Theorem 4.1, then a reversible analogue of Theorem 5.2 holds, where the terms "symplectic" and "Hamiltonian" are simply replaced by "reversible". The proof remains the same, working with the reversible analogues of the results used for the Hamiltonian case.

XII.6 Exercises

1. In the situation of the invariant manifold theorem, Theorem 3.1, suppose in addition that f and g are α-periodic in x: $f(x+\alpha,y) = f(x,y)$, $g(x+\alpha,y) = g(x,y)$ for all $x \in X$, $y \in Y$. Show that in this case the function s defining the invariant manifold is also α-periodic.
   Hint. The Hadamard transform maps α-periodic functions to α-periodic functions.
2. Show that if the time-τ flow map $\Phi = \varphi_\tau$ of a differential equation has an attractive invariant manifold $\mathcal{M}$, and if the flow φ_t maps a domain of attractivity of $\mathcal{M}$ under Φ into itself for every real t, then $\mathcal{M}$ is also invariant under the flow φ_t for every real t.
   Hint. Write $\varphi_t = \Phi^n \circ \varphi_t \circ \Phi^{-n}$ and use the attractivity of $\mathcal{M}$ for $n \to \infty$.

3. Prove that in the situation of Theorem 3.1, iterates $(x_{n+1}, y_{n+1}) = \Phi(x_n, y_n)$ have the *property of asymptotic phase* (Nipp & Stoffer 1992): there exists a sequence $(\widetilde{x}_n, \widetilde{y}_n)$ of iterates on the invariant manifold, i.e., with $(\widetilde{x}_{n+1}, \widetilde{y}_{n+1}) = \Phi(\widetilde{x}_n, \widetilde{y}_n)$ and $\widetilde{y}_n = s(\widetilde{x}_n)$, such that for all $n \geq 0$,

$$\|x_n - \widetilde{x}_n\| \leq c\,\|y_n - s(x_n)\|$$
$$\|y_n - \widetilde{y}_n\| \leq (1 + \lambda c)\,\|y_n - s(x_n)\|\ ,$$

   where $c = \lambda/(1 - \lambda\lambda^*)$ with $\lambda = 2L_{yx}/(1 - L_{xx} - L_{yy})$ of (3.5) and $\lambda^* = 2L_{xy}/(1 - L_{xx} - L_{yy})$. Note that $\|y_n - s(x_n)\| \leq \rho^n \|y_0 - s(x_0)\|$ by Theorem 3.1.
   Hint. Consider the sequences $(\widetilde{x}_n^{(k)}, \widetilde{y}_n^{(k)})$ defined by $\widetilde{x}_k^{(k)} = x_k$, $\widetilde{y}_k^{(k)} = s(x_k)$ and $(\widetilde{x}_{n+1}^{(k)}, \widetilde{y}_{n+1}^{(k)}) = \Phi(\widetilde{x}_n^{(k)}, \widetilde{y}_n^{(k)})$ for $n = k-1, \ldots, 1, 0$. Show that, for fixed n, the sequence $(x_n^{(k)})$ $(k \geq n)$ is a Cauchy sequence.
4. Show that Lemma 2.1 holds unchanged if the diophantine condition (X.2.4) for $\omega(a^*)$ is weakened to $\omega(a^*) = \omega^* + \mathcal{O}(\delta^2)$ with ω^* satisfying (X.2.4).
5. In the situation of Theorem 5.2, show that every numerical integrator of order p has an attractive invariant torus if $h^p \ll \varepsilon$. This torus is $\mathcal{O}(h^p/\varepsilon)$ close to the invariant torus of the continuous system.

Chapter XIII. Oscillatory Differential Equations with Constant High Frequencies

This chapter deals with numerical methods for second-order differential equations with oscillatory solutions. These methods are designed to require a new complete function evaluation only after a time step over one or many periods of the fastest oscillations in the system. Various such methods have been proposed in the literature – some of them decades ago, some very recently, motivated by problems from molecular dynamics, astrophysics and nonlinear wave equations. For these methods it is not obvious what implications geometric properties like symplecticity or reversibility have on the long-time behaviour, e.g., on energy conservation. The backward error analysis of Chap. IX, which was the backbone of the results of the three preceding chapters, is no longer applicable when the product of the step size with the highest frequency is not small, which is the situation of interest here. The "exponentially small" remainder terms are now only $\mathcal{O}(1)$! For differential equations where the high frequencies of the oscillations remain nearly constant along the solution, a substitute for the backward error analysis of Chap. IX is given by the *modulated Fourier expansions* of the exact and the numerical solutions. Among other properties, they permit us to understand the numerical long-time conservation of the total and oscillatory energies (or the failure of conserving energy in certain cases). It turns out, symmetry of the methods is still essential, but symplecticity plays no role in the analysis and in the numerical experiments, and new conditions of an apparently non-geometric nature come into play.

XIII.1 Towards Longer Time Steps in Solving Oscillatory Equations of Motion

> Dynamical systems with multiple time scales pose a major problem in simulations because the small time steps required for stable integration of the fast motions lead to large numbers of time steps required for the observation of slow degrees of freedom and thus to the need to compute a large number of forces.
>
> (M. Tuckerman, B.J. Berne & G.J. Martyna 1992)

We describe numerical methods that have been proposed for solving highly oscillatory second-order differential equations with fewer force evaluations than are needed by standard integrators like the Störmer–Verlet method. We present the ideas

underlying the construction of the methods and leave numerical comparisons to Sect. XIII.2 and the analysis of the methods to Sections XIII.3–XIII.6. We consider only methods that are symmetric or symplectic. The presentation in this section follows roughly the chronological order.

XIII.1.1 The Störmer–Verlet Method vs. Multiple Time Scales

> Perhaps the most widely used method of integrating the equations of motion is that initially adopted by Verlet (1967) and attributed to Störmer.
> (M.P. Allen & D.J. Tildesley 1987, p. 78)

The Newtonian equations of motion of particle systems (in molecular dynamics, astrophysics and elsewhere) are second-order differential equations

$$\ddot{q} = -\nabla V(q) . \tag{1.1}$$

To simplify the presentation, we omit the positive definite mass matrix M which would usually multiply $\ddot{q}$. This entails no loss of generality, since a transformation $q \to M^{1/2}q$ and $V(q) \to V(M^{-1/2}q)$ gives the very form (1.1).

The standard numerical integrator of molecular dynamics is the Störmer–Verlet scheme; see Chap. I. We recall that this method computes the new positions q_{n+1} at time t_{n+1} from

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f_n \tag{1.2}$$

with the force $f_n = -\nabla V(q_n)$. Velocity approximations are given by

$$\dot{q}_n = \frac{q_{n+1} - q_{n-1}}{2h} .$$

In its one-step formulation (see (I.1.17)) the method reads[1]

$$\begin{aligned} p_{n+1/2} &= p_n + \tfrac{1}{2}hf_n \\ q_{n+1} &= q_n + hp_{n+1/2} \\ p_{n+1} &= p_{n+1/2} + \tfrac{1}{2}hf_{n+1} . \end{aligned} \tag{1.3}$$

We recall that this is a symmetric and symplectic method of order 2. For linear stability, i.e., for bounded error propagation in linearized equations, the step size must be restricted to

$$h\omega < 2$$

where ω is the largest eigenfrequency (i.e., square root of an eigenvalue) of the Hessian matrix $\nabla^2 V(q)$ along the numerical solution; see Sect. I.5.1. Good energy conservation requires an even stronger restriction on the step size. Values of $h\omega \approx \frac{1}{2}$ are frequently used in molecular dynamics simulations.

The potential $V(q)$ is often a sum of potentials that act on different time scales,

[1] We write p when the Hamiltonian structure and symplecticity are an issue, and $\dot{q}$ otherwise.

$$V(q) = W(q) + U(q) \quad \text{with} \quad \begin{array}{l} \nabla^2 W(q) \text{ positive semi-definite and} \\ \|\nabla^2 W(q)\| \gg \|\nabla^2 U(q)\| \,. \end{array} \tag{1.4}$$

In this situation, solutions are in general highly oscillatory on the slow time scale $\tau \sim 1/\|\nabla^2 U(q)\|^{1/2}$.

In particular when the *fast* forces $-\nabla W(q)$ are cheaper to evaluate than the *slow* forces $-\nabla U(q)$, it is of interest to devise methods where the required number of slow-force evaluations is not (or not severely) affected by the presence of the fast forces which are responsible for the oscillatory behaviour and which restrict the step size of standard integrators like the Störmer–Verlet scheme. This situation occurs in molecular dynamics, where $W(q)$ corresponds to short-range molecular bonds, whereas $U(q)$ includes *inter alia* long-range electrostatic potentials.

In some approaches to this computational problem, the differential model is modified: highly oscillatory components are replaced by constraints (Ryckaert, Ciccotti & Berendsen 1977), or stochastic and dissipative terms are added to the model (see Schlick 1999). Such modifications may prove highly successful in some applications. In the following, however, we restrict our attention to methods which aim at long time steps directly for the problem (1.1) with (1.4).

Spatial semi-discretizations of nonlinear wave equations, such as the sine-Gordon equation

$$u_{tt} = u_{xx} - \sin u \,,$$

form another important class of equations (1.1) with (1.4). Here $W(q) = \frac{1}{2} q^T A q$, where A is the discretization matrix of the differential operator $-\partial^2/\partial x^2$.

XIII.1.2 Gautschi's and Deuflhard's Trigonometric Methods

> It is anticipated that trigonometric methods can be applied, with similar success, also to nonlinear differential equations describing oscillation phenomena. (W. Gautschi 1961)

The oldest methods allowing the use of long time steps in oscillatory problems concern the particular case of a quadratic potential $W(q) = \frac{1}{2}\omega^2 q^T q$ with $\omega \gg 1$, for which the equations take the form

$$\ddot{q} \;=\; -\,\omega^2 q + g(q) \,. \tag{1.5}$$

For such equations, Gautschi (1961) proposed a number of methods of multistep type which are constructed to be exact if the solution is a trigonometric polynomial in ωt of a prescribed degree. The simplest of these methods (and the only symmetric one) reads

$$q_{n+1} - 2q_n + q_{n-1} \;=\; h^2 \,\mathrm{sinc}^2(\tfrac{1}{2}h\omega)\, \ddot{q}_n \,, \tag{1.6}$$

where $\mathrm{sinc}\,\xi = \sin\xi/\xi$ and $\ddot{q}_n = -\omega^2 q_n + g_n$ with $g_n = g(q_n)$, or equivalently

$$q_{n+1} - 2\cos(h\omega) q_n + q_{n-1} \;=\; h^2 \,\mathrm{sinc}^2(\tfrac{1}{2}h\omega)\, g_n \,. \tag{1.7}$$

The method gives the exact solution for equations (1.5) with $g = Const$ and arbitrary ω (see also Hersch (1958) for such a construction principle). This property is readily verified with the variation-of-constants formula

$$\begin{aligned}\begin{pmatrix} q(t) \\ \dot q(t) \end{pmatrix} &= \begin{pmatrix} \cos t\omega & \omega^{-1}\sin t\omega \\ -\omega \sin t\omega & \cos t\omega \end{pmatrix} \begin{pmatrix} q_0 \\ \dot q_0 \end{pmatrix} \\ &\quad + \int_0^t \begin{pmatrix} \omega^{-1}\sin(t-s)\omega \\ \cos(t-s)\omega \end{pmatrix} g\big(q(s)\big)\, ds\ . \end{aligned} \tag{1.8}$$

This formula also shows that the following scheme for a velocity approximation becomes exact for $g = Const$:

$$\dot q_{n+1} - \dot q_{n-1} = 2h\,\mathrm{sinc}(h\omega)\,\ddot q_n\ . \tag{1.9}$$

Starting values q_1 and $\dot q_1$ are also obtained from (1.8) with $g(q_0)$ in place of $g(q(s))$.

Deuflhard (1979) considered h^2-extrapolation based on the explicit symmetric method that is obtained by replacing the integral term in (1.8) by its trapezoidal rule approximation:

$$\begin{pmatrix} q_{n+1} \\ h\dot q_{n+1} \end{pmatrix} = \begin{pmatrix} \cos h\omega & \mathrm{sinc}\, h\omega \\ -h\omega\sin h\omega & \cos h\omega \end{pmatrix} \begin{pmatrix} q_n \\ h\dot q_n \end{pmatrix} + \frac{h^2}{2} \begin{pmatrix} \mathrm{sinc}(h\omega)\, g_n \\ g_{n+1} + \cos(h\omega)\, g_n \end{pmatrix}. \tag{1.10}$$

Eliminating the velocities yields the two-step formulation

$$q_{n+1} - 2\cos(h\omega)q_n + q_{n-1} = h^2\,\mathrm{sinc}(h\omega)\, g_n\ . \tag{1.11}$$

The velocity approximation is obtained back from

$$2h\,\mathrm{sinc}(h\omega)\,\dot q_n = q_{n+1} - q_{n-1} \tag{1.12}$$

or alternatively from

$$\dot q_{n+1} - 2\cos(h\omega)\dot q_n + \dot q_{n-1} = h^2\, \frac{g_{n+1} - g_{n-1}}{2h}\ .$$

Both Gautschi's and Deuflhard's method reduce to the Störmer–Verlet scheme for $\omega = 0$. Both methods extend in a straightforward way to systems

$$\ddot q = -Aq + g(q) \tag{1.13}$$

with a symmetric positive semi-definite matrix A, by formally replacing ω by $\Omega = A^{1/2}$ in the above formulas. The methods then require the computation of products of entire functions of the matrix $h^2 A$ with vectors. This can be done by diagonalizing A, which is efficient for problems of small dimension or in spectral methods for nonlinear wave equations. In high-dimensional problems where a diagonalization is not feasible, these matrix function times vector products can be efficiently computed by superlinearly convergent Krylov subspace methods, see Druskin & Knizhnerman (1995) and Hochbruck & Lubich (1997).

The above methods permit extensions to more general problems (1.1) with (1.4), but this requires a reinterpretation to which we turn next.

XIII.1.3 The Impulse Method

> Integrators based on r-RESPA [...] have led to considerable speed-up in the CPU time for large scale simulations of biomacromolecular solutions. Since r-RESPA is symplectic such integrators are very stable.
>
> (B.J. Berne 1999)

The Störmer–Verlet method (1.3) can be interpreted as approximating the flow φ_h^H of the system with Hamiltonian $H(p,q) = T(p) + V(q)$ with $T(p) = \frac{1}{2}p^T p$ by the symmetric splitting

$$\varphi_{h/2}^V \circ \varphi_h^T \circ \varphi_{h/2}^V ,$$

which involves only the flows of the systems with Hamiltonians $T(p)$ and $V(q)$, which are trivial to compute; see Sect. II.5.

In the situation (1.4) of a potential $V = W + U$, we may instead use a different splitting of $H = (T + W) + U$ and approximate the flow φ_h^H of the system by

$$\varphi_{h/2}^U \circ \varphi_h^{T+W} \circ \varphi_{h/2}^U .$$

This gives a method that was proposed in the context of molecular dynamics by Grubmüller, Heller, Windemuth & Schulten (1991) (their Verlet-I scheme) and by Tuckerman, Berne & Martyna (1992) (their r-RESPA scheme). Following the terminology of García-Archilla, Sanz-Serna & Skeel (1999) we here refer to this method as the *impulse method*:

$$\begin{array}{ll} \text{1. kick: set} & p_n^+ = p_n - \frac{1}{2}h\,\nabla U(q_n) \\ \text{2. oscillate:} & \text{solve } \ddot q = -\nabla W(q) \text{ with initial values } (q_n, p_n^+) \\ & \text{over a time step } h \text{ to obtain } (q_{n+1}, p_{n+1}^-) \\ \text{3. kick: set} & p_{n+1} = p_{n+1}^- - \frac{1}{2}h\nabla U(q_{n+1}) \end{array} \tag{1.14}$$

Step 2 must in general be computed approximately by a numerical integrator with a smaller time step, which results in the multiple time stepping method that we encountered in Sect. VIII.4. If the inner integrator is symplectic and symmetric, as it would be for the natural choice of the Störmer–Verlet method, then also the overall method is symplectic – as a composition of symplectic transformations, and it is symmetric – as a symmetric composition of symmetric steps.

It is interesting to note that the impulse method (with exact solution of step 2) reduces to Deuflhard's method in the case of a quadratic potential $W(q) = \frac{1}{2}q^T Aq$ (Exercise 1).

Though the method does allow larger step sizes than the Störmer–Verlet method in molecular dynamics simulations, it is not free from numerical difficulties. Biesadecki & Skeel (1993) and García-Archilla et al. (1999) report and in linear model problems analyze instabilities and numerical resonance phenomena when the product of the step size h with an eigenfrequency ω of $\nabla^2 W$ is near an integral multiple of π.

XIII.1.4 The Mollified Impulse Method

> We also propose a nontrivial improvement of the impulse method that we call the *mollified impulse method*, for which superior stability and accuracy is demonstrated.
>
> (B.García-Archilla, J.M. Sanz-Serna & R.D. Skeel 1999)

Difficulties with the impulse method can be intuitively seen to come from two sources: the slow force $-\nabla U(q)$ has an effect only at the ends of a time step, but it does not enter into the oscillations in between; the slow force is evaluated, somewhat arbitrarily, at isolated points of the oscillatory solution.

García-Archilla et al. (1999) propose to evaluate the slow force at an *averaged* value $\overline{q}_n = a(q_n)$. They replace the potential $U(q)$ by $\overline{U}(q) = U(a(q))$ and hence the slow force $-\nabla U(q)$ in the impulse method by the *mollified force*

$$-\nabla\overline{U}(q) = -a'(q)^T\nabla U(a(q)) \, . \tag{1.15}$$

Since this *mollified impulse method* is the impulse method for a modified potential, it is again symplectic and symmetric.

There are numerous possibilities to choose the average $a(q_n)$, but care should be taken that it is only a function of the position q_n and thus independent of p_n, in order to obtain a symplectic and symmetric method. This precludes taking averages of the solution of the problem in the oscillation step (Step 2) of the algorithm. Instead, one solves the auxiliary initial value problem

$$\ddot{x} = -\nabla W(x) \quad \text{with} \quad x(0) = q,\ \dot{x}(0) = 0 \tag{1.16}$$

together with the variational equation (using the same method and the same step size)

$$\ddot{X} = -\nabla^2 W(x(t))X \quad \text{with} \quad X(0) = I,\ \dot{X}(0) = 0 \tag{1.17}$$

and computes the time average over an interval of length ch for some $c > 0$:

$$a(q) = \frac{1}{ch}\int_0^{ch} x(t)\,dt \, , \quad a'(q) = \frac{1}{ch}\int_0^{ch} X(t)\,dt \, . \tag{1.18}$$

García-Archilla et al. (1999) found that the choice $c = 1$ gives the best results. Weighted averages instead of the simple average used above give no improvement.

Izaguirre, Reich & Skeel (1999) propose to take $a(q)$ as a projection of q to the manifold $\nabla W(q) = 0$ of rest positions of the fast forces, for situations where all non-zero eigenfrequencies of $\nabla^2 W(q)$ are much larger than those of $\nabla^2 U(q)$. This choice is motivated by the fact that solutions oscillate about this manifold.

We now turn to the interesting special case of a quadratic $W(q) = \frac{1}{2}q^T Aq$ with a symmetric positive semi-definite matrix A. In this case, the above average can be computed analytically. It becomes

$$a(q) = \phi(h\Omega)q$$

with $\Omega = A^{1/2}$ and the function $\phi(\xi) = \mathrm{sinc}(c\xi)$. For $a(q)$ defined by the orthogonal projection to $Aq = 0$ we have $\phi(0) = 1$ and $\phi(\xi) = 0$ for ξ away from 0. With $g_n = -\phi(h\Omega)\nabla U(\phi(h\Omega)q_n)$, the mollified impulse method reduces to

$$\begin{aligned} p_n^+ &= p_n + \tfrac{1}{2}hg_n \\ \begin{pmatrix} q_{n+1} \\ p_{n+1}^- \end{pmatrix} &= \begin{pmatrix} \cos h\Omega & h\,\mathrm{sinc}\, h\Omega \\ -\Omega \sin h\Omega & \cos h\Omega \end{pmatrix} \begin{pmatrix} q_n \\ p_n^+ \end{pmatrix} \\ p_{n+1} &= p_{n+1}^- + \tfrac{1}{2}hg_{n+1}\,. \end{aligned} \tag{1.19}$$

This can equivalently be written as (1.10) with the same g_n (and Ω in place of ω), or in the two-step form (1.11) with (1.12).

XIII.1.5 Gautschi's Method Revisited

We recall that Gautschi's method (1.7) (with $\Omega = A^{1/2}$ in place of ω) integrates equations $\ddot{q} = -Aq + g(q)$ exactly in the case of a constant inhomogeneity $g(q) = Const$. This property is obviously kept if the argument of g in the algorithm is modified to

$$g_n = g(\phi(h\Omega)q_n)$$

similar to the previous subsection. Such Gautschi-type methods were analyzed by Hochbruck & Lubich (1999a). Functions ϕ with $\phi(0) = 1$ that vanish at integral multiples of π give a substantial improvement over the original Gautschi method. The choice

$$\phi(\xi) = \mathrm{sinc}\,\xi\big(1 + \tfrac{1}{3}\sin^2 \tfrac{1}{2}\xi\big) \tag{1.20}$$

was found to give particularly good accuracy. The methods are symmetric but not symplectic.

The following symmetric method for general problems (1.1) with (1.4) was proposed by Hochbruck & Lubich (1999a). The method reduces to Gautschi-type methods for quadratic $W(q) = \frac{1}{2}q^T Aq$. Given q_n and $\dot{q}_n$, one computes an averaged value $\overline{q}_n = a(q_n)$ and the solution of

$$\ddot{u} = -\nabla W(u) - \nabla U(\overline{q}_n) \qquad \text{with} \quad u(0) = q_n,\ \dot{u}(0) = \dot{q}_n \tag{1.21}$$

backwards and forwards on the intervals from 0 to $-h$ and 0 to h. Note that this requires only one evaluation of the slow force $-\nabla U$. Then, q_{n+1} and $\dot{q}_{n+1}$ are computed from

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= u(h) - 2u(0) + u(-h) \\ \dot{q}_{n+1} - \dot{q}_{n-1} &= \dot{u}(h) - \dot{u}(-h)\,. \end{aligned} \tag{1.22}$$

When the differential equation for u is solved approximately by a symmetric numerical method with smaller time steps, then this becomes a symmetric multiple time-stepping method. For the interpretation as an averaged-force method and for the corresponding one-step version, where the initial value for the velocity in (1.21) is replaced by $\dot{u}(0) = 0$, we refer back to Sect. VIII.4 (where q_n instead of the average $\overline{q}_n = a(q_n)$ was taken as the argument of the slow force $-\nabla U$).

XIII.1.6 Two-Force Methods

Hairer & Lubich (2000a) compare the analytical solution and the numerical solutions given by the above methods in the Fermi–Pasta–Ulam model of Sect. I.5.1, using the tool of modulated Fourier expansions (see Sections XIII.3 and XIII.5 below). Their analysis of the slow energy exchange between stiff springs leads them to propose the following method for equations $\ddot{q} = -Aq + g(q)$, which requires two evaluations of the slow force per time step: with $\Omega = A^{1/2}$, set

$$q_{n+1} - 2\cos(h\Omega)\, q_n + q_{n-1} \;=\; h^2 \operatorname{sinc}(h\Omega)\, g(q_n) + h^2 d_n \tag{1.23}$$

with

$$d_n \;=\; \operatorname{sinc}^2(h\Omega)\, g(q_n) - \operatorname{sinc}(h\Omega)\, g\big(\operatorname{sinc}(h\Omega) q_n\big)\, . \tag{1.24}$$

This method gives the correct slow energy exchange between stiff components in the model problem and has better energy conservation than the Deuflhard/impulse method. With the velocity approximation (1.12) the method can equivalently be written in the one-step forms (1.19) or (1.10). The method extends again to a symmetric method for general problems (1.1) with (1.4), giving a correction to the impulse method: let $g(q) = -\nabla U(q)$ and let $a(q)$ be defined by (1.18) with $c = 1$. Set $\overline{q}_n = a(q_n)$ and

$$\overline{g}(q_n) \;=\; \frac{2}{h^2}\Big(a\big(q_n + \tfrac{1}{2}h^2 g(q_n)\big) - a(q_n) \Big)\, .$$

The method then consists of taking

$$g_n \;=\; g(q_n) + \overline{g}(q_n) - g(\overline{q}_n)$$

instead of $g(q_n) = -\nabla U(q_n)$ in the impulse method (1.14).

A two-force method with interesting properties, for situations where all non-zero eigenfrequencies of A are much larger than those of $\nabla^2 U(q)$, is given by (1.23) with

$$d_n \;=\; \operatorname{sinc}^2(\tfrac{1}{2}h\Omega)\, g\big(\chi(h\Omega) q_n\big) - \operatorname{sinc}(h\Omega)\, g\big(\chi(h\Omega) q_n\big)\, , \tag{1.25}$$

where $\chi(0) = 1$ and $\chi(\xi) = 0$ for ξ away from 0.

XIII.2 A Nonlinear Model Problem and Numerical Phenomena

To gain insight into the properties of the various numerical methods described in the previous section, it is helpful to study the methods when they are applied to suitably chosen, rather simple model problems which show characteristic features but are still accessible to an analysis. Such an approach has traditionally been very successful for stiff differential equations (see, e.g., Hairer & Wanner 1996). For the

present stiff-oscillatory case we investigate the behaviour of the numerical methods on nonlinear systems

$$\ddot{x} + \Omega^2 x = g(x) \tag{2.1}$$

with a smooth gradient nonlinearity $g(x) = -\nabla U(x)$ and with the square matrix

$$\Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix} , \quad \omega \gg 1 , \tag{2.2}$$

with blocks of arbitrary dimension. We consider only solutions whose energy is bounded independently of ω, so that in particular the initial values satisfy

$$\tfrac{1}{2}\|\dot{x}(0)\|^2 + \tfrac{1}{2}\|\Omega x(0)\|^2 \le E \tag{2.3}$$

with E independent of ω.

The Fermi–Pasta–Ulam (FPU) problem of Sect. I.5.1 belongs precisely to this class, and we will present numerical experiments with this example. In the model problem (2.1) with (2.2) we clearly impose strong restrictions in that the high frequencies are confined to the linear part and that there is a single, constant high frequency. The extension to several high frequencies will be given in Sect. XIII.9, and constant-frequency systems with a position-dependent kinetic energy term are considered in Sect. XIII.10. Oscillatory systems with time- or solution-dependent high frequencies will be studied, with different techniques and for different numerical methods, in Chap. XIV.

In any case, satisfactory behaviour of a method on the model problem (2.1) can be anticipated to be necessary for a successful treatment of more general situations.

XIII.2.1 Time Scales in the Fermi–Pasta–Ulam Problem

The FPU model shows different behaviour on different time scales: almost-harmonic motion of the stiff springs on the time scale ω^{-1}, motion of the soft springs on the scale ω^0, energy exchange between stiff springs on the time scale ω, and almost-preservation of the oscillatory energy over intervals that are exponentially long in ω. This is illustrated in the following.

We consider the FPU problem with three stiff springs with the data of Sect. I.5.1. The four pictures of Fig. 2.1 show the evolution of the following quantities: the total energy

$$H(x, \dot{x}) = \tfrac{1}{2}\dot{x}^T \dot{x} + \tfrac{1}{2}x^T \Omega^2 x + U(x) , \tag{2.4}$$

(or rather $H - 0.8$ for graphical reasons), which is a conserved quantity; the oscillatory energy

$$I = I_1 + I_2 + I_3 \quad \text{with} \quad I_j = \tfrac{1}{2}\, \dot{x}_{1,j}^2 + \tfrac{1}{2}\, \omega^2\, x_{1,j}^2 , \tag{2.5}$$

where $x_{1,j}$ is the jth component of the lower half $x_1 \in \mathbb{R}^3$ of $x = (x_0, x_1)^T \in \mathbb{R}^6$, decomposed according to the blocks of Ω in (2.2). We recall that $x_{1,j}$ represents the

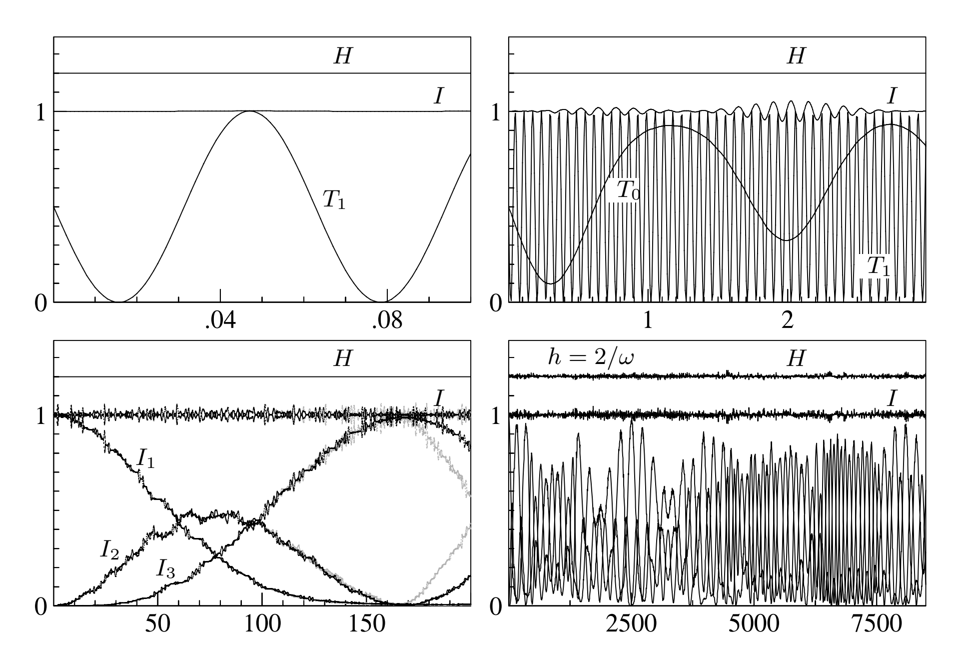


Fig. 2.1. Different time scales in the Fermi–Pasta–Ulam model ($\omega = 50$)

elongation of the jth stiff spring. Further quantities shown are the kinetic energy of the mass centre motion and of the relative motion of masses joined by a stiff spring,

$$T_0 = \tfrac{1}{2}\|\dot{x}_0\|^2 , \qquad T_1 = \tfrac{1}{2}\|\dot{x}_1\|^2 .$$

Time Scale ω^{-1}. The vibration of the stiff linear springs is nearly harmonic with almost-period π/ω. This is illustrated by the plot of T_1 in the first picture.

Time Scale ω^0. This is the time scale of the motion of the soft nonlinear springs, as is exemplified by the plot of T_0 in the second picture of Fig. 2.1.

Time Scale ω. A slow energy exchange among the stiff springs takes place on the scale ω. In the third picture, the initially excited first stiff spring passes energy to the second one, and then also the third stiff spring begins to vibrate. The picture also illustrates that the problem is very sensitive to perturbations of the initial data: the grey curves of each of I_1, I_2, I_3 correspond to initial data where 10^{-5} has been added to $x_{0,1}(0)$, $\dot{x}_{0,1}(0)$ and $\dot{x}_{1,1}(0)$. The displayed solutions of the first three pictures have been computed very accurately by an adaptive integrator.

Time Scale ω^N, $N \geq 2$. The oscillatory energy I has only $\mathcal{O}(\omega^{-1})$ deviations from the initial value over very long time intervals. The fourth picture of Fig. 2.1 shows the total energy H and the oscillatory energy I as computed by method (1.10)-(1.11) of Sect. XIII.1.2 with the step size $h = 2/\omega$, which is nearly as large as the length of the time interval of the first picture. No drift is seen for H or I.

XIII.2.2 Numerical Methods

The methods described in Sect. XIII.1 all have in common that they reduce to the Störmer–Verlet method when they are applied to (2.1) with $\Omega = 0$, and they become exact solvers for the linear homogeneous problem with $g(x) \equiv 0$. They can be formulated as one-step or two-step schemes.

Two-Step Formulation. All the methods of Sections XIII.1.2–XIII.1.5, when applied to the system (2.1), can be written in the two-step form

$$x_{n+1} - 2\cos(h\Omega)\, x_n + x_{n-1} \;=\; h^2 \Psi g(\Phi x_n)\ . \tag{2.6}$$

Here $\Psi = \psi(h\Omega)$ and $\Phi = \phi(h\Omega)$, where the *filter functions* ψ and ϕ are even, real-valued functions with $\psi(0) = \phi(0) = 1$. In our numerical experiments we will consider the following choices of ψ and ϕ, where again $\operatorname{sinc}(\xi) = \sin\xi/\xi$:

(A)	$\psi(\xi) = \operatorname{sinc}^2(\frac12\xi)$	$\phi(\xi) = 1$	Gautschi (1961)
(B)	$\psi(\xi) = \operatorname{sinc}(\xi)$	$\phi(\xi) = 1$	Deuflhard (1979)
(C)	$\psi(\xi) = \operatorname{sinc}(\xi)\,\phi(\xi)$	$\phi(\xi) = \operatorname{sinc}(\xi)$	García-Archilla & al. (1999)
(D)	$\psi(\xi) = \operatorname{sinc}^2(\frac12\xi)$	$\phi(\xi)$ of (1.20)	Hochbruck & Lubich (1999a)
(E)	$\psi(\xi) = \operatorname{sinc}^2(\xi)$	$\phi(\xi) = 1$	Hairer & Lubich (2000a)

One-Step Formulation. The method (2.6) can be written as a symmetric one-step method of a form that is motivated by the variation-of-constants formula (1.8). This now also includes a velocity approximation $\dot x_n$:

$$x_{n+1} = \cos h\Omega\, x_n + \Omega^{-1}\sin h\Omega\, \dot x_n + \tfrac12 h^2 \Psi\, g_n \tag{2.7}$$

$$\dot x_{n+1} = -\Omega\sin h\Omega\, x_n + \cos h\Omega\, \dot x_n + \tfrac12 h\big(\Psi_0\, g_n + \Psi_1\, g_{n+1}\big) \tag{2.8}$$

where $g_n = g(\Phi x_n)$ and $\Psi_0 = \psi_0(h\Omega)$, $\Psi_1 = \psi_1(h\Omega)$ with even functions ψ_0, ψ_1 satisfying $\psi_0(0) = 1$, $\psi_1(0) = 1$. Exchanging $n \leftrightarrow n+1$ and $h \leftrightarrow -h$ in the method, it is seen that the method is symmetric if and only if

$$\psi(\xi) = \operatorname{sinc}(\xi)\,\psi_1(\xi)\,, \qquad \psi_0(\xi) = \cos(\xi)\,\psi_1(\xi)\ . \tag{2.9}$$

The method is then symplectic if and only if (Exercise 2)

$$\psi(\xi) = \operatorname{sinc}(\xi)\,\phi(\xi)\ . \tag{2.10}$$

Two-Step Velocity Schemes. For a symmetric method (2.7)–(2.8) the velocity approximation can be equivalently obtained from

$$2h\operatorname{sinc}(h\omega)\,\dot x_n = x_{n+1} - x_{n-1} \tag{2.11}$$

(for $\sin(h\omega) \neq 0$) or from

$$\dot x_{n+1} - 2\cos(h\Omega)\,\dot x_n + \dot x_{n-1} \;=\; \tfrac12 h\,\Psi_1(g_{n+1} - g_{n-1})\ . \tag{2.12}$$

The latter formula gives a symmetric two-step method for arbitrary even functions ψ_1 with $\psi_1(0) = 1$, which do not necessarily satisfy (2.9).

Multi-Force Methods. The methods of Sect. XIII.1.6 belong to the class of multi-force methods, which generalize the right-hand side of (2.6) to a linear combination of such terms:

$$x_{n+1} - 2\cos(h\Omega)\, x_n + x_{n-1} \;=\; h^2 \sum_{j=1}^{k} \Psi_j \, g(\Phi_j x_n) \tag{2.13}$$

with $\Psi_j = \psi_j(h\Omega)$, $\Phi_j = \phi_j(h\Omega)$, where ψ_j, ϕ_j are even functions with

$$\sum_{j=1}^{k} \psi_j(0) = 1\,, \qquad \phi_j(0) = 1 \qquad \text{for } j = 1, \ldots, k\,.$$

In our numerical experiments we include the method

(F) two-force method (1.23) with (1.24).

XIII.2.3 Accuracy Comparisons

The accuracy of the methods (A)-(E) and the Störmer–Verlet method on a short time interval is shown in Fig. 2.2, where the errors at $t = 1$ of the different solution components in the FPU problem (with $\omega = 50$) are plotted as a function of the step size h. Here and in all the following numerical experiments, the methods were

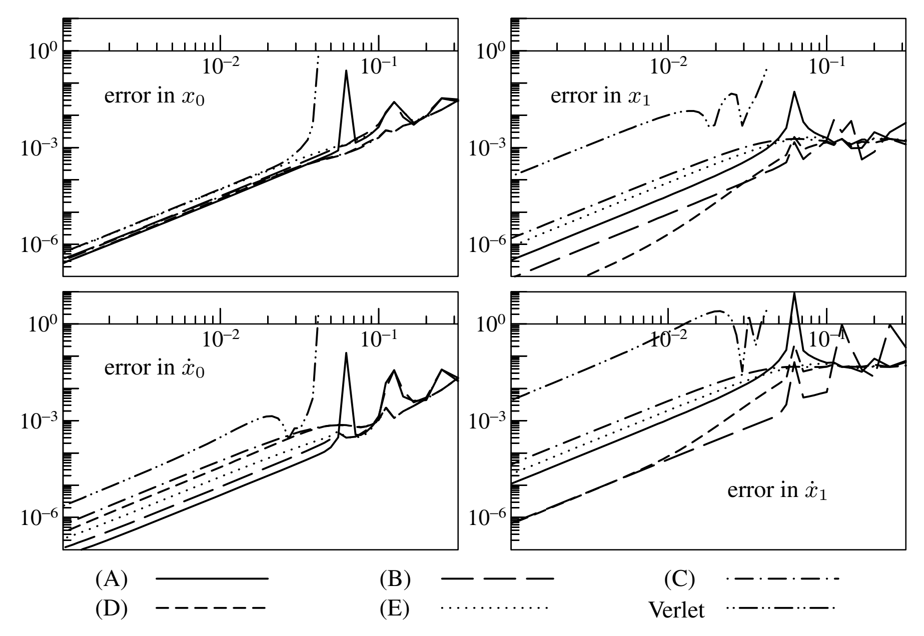


Fig. 2.2. Global error at $t = 1$ for the different components and for the five methods (A) - (E) and the Störmer–Verlet method as a function of the step size h

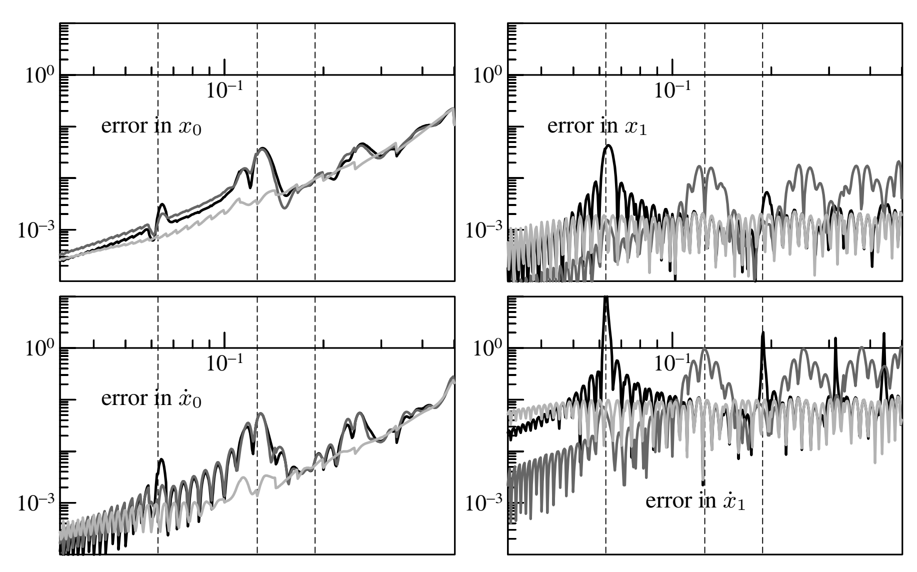


Fig. 2.3. Global error at the first grid point after $t = 1$ for the different components as a function of the step size h. The error for method (A) is drawn in black, for method (B) in dark grey, and for method (C) in light grey. The vertical lines indicate step sizes for which $h\omega$ equals π, 2π, or 3π

implemented in the one-step formulation (2.7)-(2.8) with (2.9). The errors in the x_0-components are nearly identical for all the methods in the stability range of the Störmer–Verlet method ($h\omega < 2$). Differences between the methods are however visible for larger step sizes. For the other solution components x_1, $\dot{x}_0$, $\dot{x}_1$ there are pronounced differences in the error behaviour of the methods. All five methods (A)-(E) are considerably more accurate than the Störmer–Verlet method. Figure 2.3 shows the errors of methods (A)-(C) for step sizes beyond the stability range of the Störmer–Verlet method. Methods (A) and (B) lose accuracy when $h\omega$ is near integral multiples of π, a phenomenon that does not occur with method (C).

XIII.2.4 Energy Exchange between Stiff Components

Figure 2.4 shows the energy exchange of the six methods (A)-(F) applied to the Fermi–Pasta–Ulam problem with the same data as in Fig. 2.1. The figures show again the oscillatory energies I_1, I_2, I_3 of the stiff springs, their sum $I = I_1 + I_2 + I_3$ and the total energy $H - 0.8$ as functions of time on the interval $0 \leq t \leq 200$. Only the methods (B), (D) and (F) give a good approximation of the energy exchange between the stiff springs. It will turn out in Sect. XIII.4.2 that a necessary condition for a correct approximation of the energy exchange is $\psi(h\omega)\phi(h\omega) = \operatorname{sinc}(h\omega)$, which is satisfied for method (B). The two-force method (F) satisfies an analogous condition for multi-force methods. The good behaviour of method (D) comes from the fact that here $\psi(h\omega)\phi(h\omega) \approx 0.95 \operatorname{sinc}(h\omega)$ for $h\omega = 1.5$.

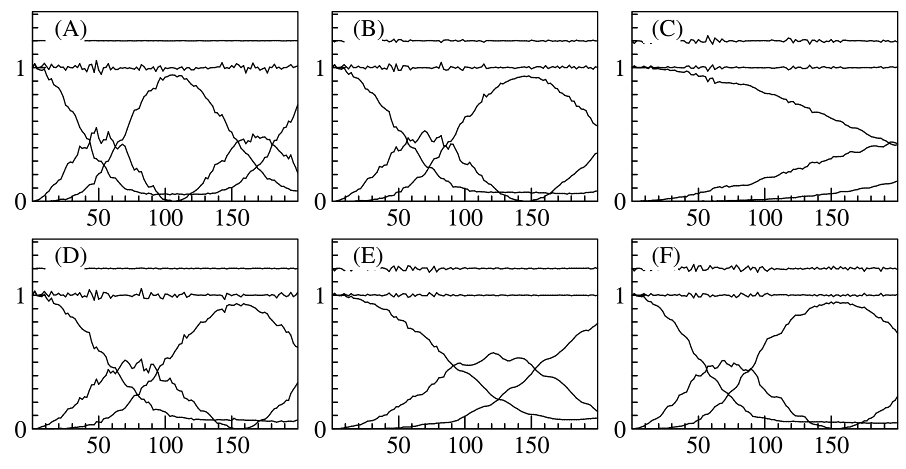


Fig. 2.4. Energy exchange between stiff springs for methods (A)-(F) ($h = 0.03$, $\omega = 50$)

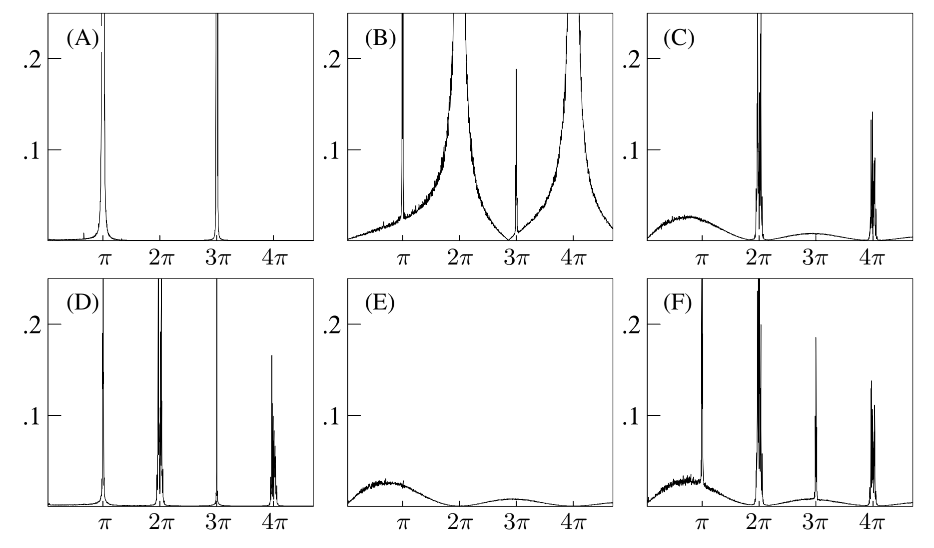


Fig. 2.5. Maximum error of the total energy on the interval $[0, 1000]$ for methods (A) - (F) as a function of $h\omega$ (step size $h = 0.02$)

XIII.2.5 Near-Conservation of Total and Oscillatory Energy

Figure 2.5 shows the maximum error of the total energy H as a function of the scaled frequency $h\omega$ (step size $h = 0.02$). We consider the long time interval $[0, 1000]$. The pictures for the different methods show that in general the total energy is well conserved. Exceptions are near integral multiples of π. Certain methods show a bad energy conservation close to odd multiples of π, other methods close to even multiples of π. Only method (E) shows a uniformly good behaviour for all frequencies. In Fig. 2.6 we show in more detail what happens close to such integral multiples of π.

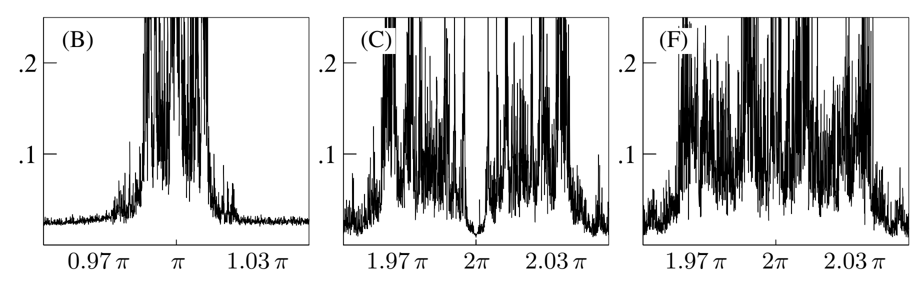


Fig. 2.6. Zoom (close to π or 2π) of the maximum error of the total energy on the interval $[0, 1000]$ for three methods as a function of $h\omega$ (step size $h = 0.02$)

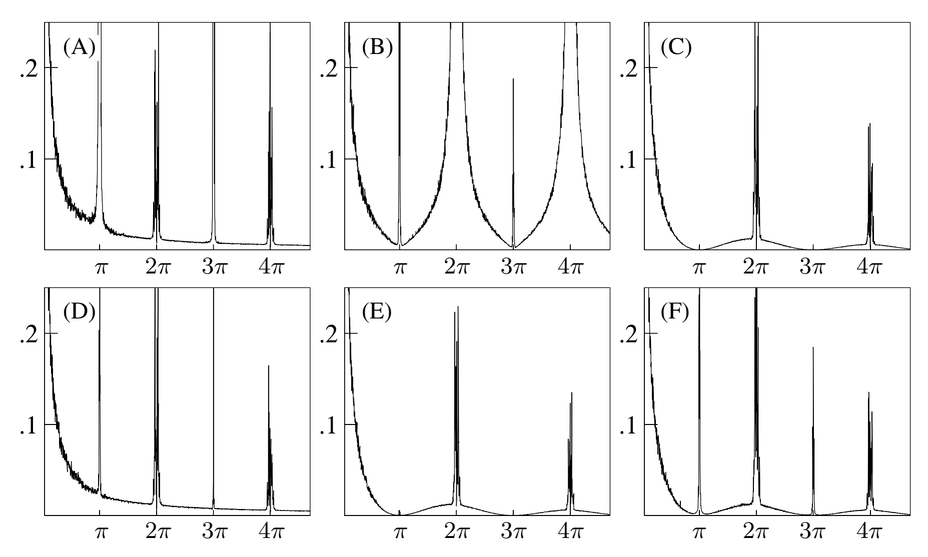


Fig. 2.7. Maximum deviation of the oscillatory energy on the interval $[0, 1000]$ for methods (A) - (F) as a function of $h\omega$ (step size $h = 0.02$)

If there is a difficulty close to π, it is typically in an entire neighbourhood. Close to 2π, the picture is different. Method (C) has good energy conservation for values of $h\omega$ that are very close to 2π, but there are small intervals to the left and to the right, where the error in the total energy is large. Unlike the other methods shown, method (B) has poor energy conservation in rather large intervals around even multiples of π. Methods (A) and (D) conserve the total energy particularly well, for $h\omega$ away from integral multiples of π.

Figure 2.7 shows similar pictures where the total energy H is replaced by the oscillatory energy I (cf. Sect. XIII.2.1). For the exact solution we have $I(t) = Const + \mathcal{O}(\omega^{-1})$. It is therefore not surprising that this quantity is not well conserved for small values of ω. For larger values of ω, we observe that the methods have difficulties in conserving the oscillatory energy when $h\omega$ is near integral multiples of π. None of the considered methods conserves both quantities H and I uniformly for all values of $h\omega$.

XIII.3 Principal Terms of the Modulated Fourier Expansion

The analytical tool for understanding the above numerical phenomena is provided by *modulated Fourier expansions*, which decompose both the exact and the numerical solution into a slowly varying part and into oscillatory components built up of trigonometric functions multiplied with slowly varying coefficient functions. A comparison of these expansions will serve as a partial substitute for the backward error analysis of Chap. IX, which yields results only for $h\omega \to 0$ and is not applicable to the situation of $h\omega \geq c > 0$ that is of interest here. In this section we derive the first terms of the modulated Fourier expansion.

XIII.3.1 Decomposition of the Exact Solution

Every solution of the linear equation $\ddot{x} + \Omega^2 x = g(t)$ with Ω of (2.2) can be written as $y(t) + \cos(\omega t)\, u(t) + \sin(\omega t)\, v(t) + \mathcal{O}(\omega^{-N})$ (for $\omega \to \infty$), where $y(t)$, $u(t)$, $v(t)$ are truncated asymptotic expansions in powers of ω^{-1} (see Exercise 4). These functions have the property that all their derivatives are bounded independently of the parameter $\omega \gg 1$. Here and in the following, a *smooth* function is understood to be a function with this property. We may hope to find a similar decomposition for solutions of the nonlinear problem (2.1). So we look for a smooth real-valued function $y(t)$ and a smooth complex-valued function $z(t) = u(t) + iv(t)$ such that the function

$$x_*(t) = y(t) + e^{i\omega t} z(t) + e^{-i\omega t} \overline{z}(t) \tag{3.1}$$

gives a small defect when it is inserted into the differential equation (2.1) and has the given initial values

$$x_*(0) = x(0)\,, \quad \dot{x}_*(0) = \dot{x}(0)\,. \tag{3.2}$$

Under the condition (2.3) the exact solution $x(t)$ has bounded energy, and we may expect the same of the approximation $x_*(t)$, which would then imply $z(t) = \mathcal{O}(\omega^{-1})$. We therefore insert the ansatz (3.1) into the differential equation (2.1) and expand the nonlinearity around the smooth part $y(t)$. With the variables $y = (y_0, y_1)$, $z = (z_0, z_1)$ partitioned according to the blocks of Ω, this gives the expressions

$$\begin{aligned} \ddot{x}_* + \Omega^2 x_* \;=\; & \begin{pmatrix} \ddot{y}_0 \\ \ddot{y}_1 + \omega^2 y_1 \end{pmatrix} \;+\; e^{i\omega t} \begin{pmatrix} -\omega^2 z_0 + 2i\omega \dot{z}_0 + \ddot{z}_0 \\ 2i\omega \dot{z}_1 + \ddot{z}_1 \end{pmatrix} \\ & \;+\; e^{-i\omega t} \begin{pmatrix} -\omega^2 \overline{z}_0 - 2i\omega \dot{\overline{z}}_0 + \ddot{\overline{z}}_0 \\ -2i\omega \dot{\overline{z}}_1 + \ddot{\overline{z}}_1 \end{pmatrix} \end{aligned}$$

and, as long as $z(t) = \mathcal{O}(\omega^{-1})$,

$$\begin{aligned} g(x_*) \;=\; & g(y) + g''(y)(z, \overline{z}) + e^{i\omega t} g'(y) z + e^{-i\omega t} g'(y) \overline{z} \\ & + e^{2i\omega t} \tfrac{1}{2} g''(y)(z, z) + e^{-2i\omega t} \tfrac{1}{2} g''(y)(\overline{z}, \overline{z}) + \mathcal{O}(\omega^{-3})\,. \end{aligned}$$

Equations for the Coefficient Functions. We now compare the coefficients of $1, e^{i\omega t}, e^{-i\omega t}$ and require that the dominant terms in these expressions be equal:

$$\begin{aligned} \ddot{y}_0 &= g_0(y) + g_0''(y)(z,\overline{z}) \\ \omega^2 y_1 &= g_1(y) \\ -\omega^2 z_0 &= g_0'(y)z \\ 2i\omega\dot{z}_1 &= g_1'(y)z\,. \end{aligned} \tag{3.3}$$

This gives a system of differential equations for y_0, z_1 and expresses y_1, z_0 as functions of y_0, z_1. We note that y_0 evolves on the time scale 1, whereas z_1 changes on the slow time scale ω. As long as $y_0(t)$ stays in a bounded domain and $z_1(t) = \mathcal{O}(\omega^{-1})$, (3.3) implies the bounds

$$y_1(t) = \mathcal{O}(\omega^{-2})\,, \quad z_0(t) = \mathcal{O}(\omega^{-3})\,, \quad \dot{z}_1(t) = \mathcal{O}(\omega^{-2})\,. \tag{3.4}$$

Initial Values. The initial values $y_0(0), \dot{y}_0(0)$ and $z_1(0)$ are obtained from condition (3.2), which gives a system that can be solved by fixed point iteration to yield

$$\begin{aligned} y_0(0) &= x_{0,0} + \mathcal{O}(\omega^{-3})\,, & \dot{y}_0(0) &= \dot{x}_{0,0} + \mathcal{O}(\omega^{-2}) \\ 2\,\mathrm{Re}\, z_1(0) &= x_{0,1} + \mathcal{O}(\omega^{-2})\,, & -\omega\, 2\,\mathrm{Im}\, z_1(0) &= \dot{x}_{0,1} + \mathcal{O}(\omega^{-2})\,. \end{aligned} \tag{3.5}$$

Defect. As long as $z_1(t) = \mathcal{O}(\omega^{-1})$, the above equations show that the defect

$$d(t) = \ddot{x}_*(t) + \Omega^2 x_*(t) - g\big(x_*(t)\big)$$

is of the form

$$d(t) = \mathrm{Re}\begin{pmatrix} \omega^{-2}e^{i\omega t}a(t) + \omega^{-2}e^{2i\omega t}b(t) + \mathcal{O}(\omega^{-3}) \\ \mathcal{O}(\omega^{-2}) \end{pmatrix} \tag{3.6}$$

with smooth functions a, b. Together with (3.3) this also shows that the smooth $\mathcal{O}(\omega^{-2})$-term $g''(y)(z,\overline{z})$ is the principal term describing the influence of oscillatory solution components on the evolution of smooth components.

Example. To illustrate the approximation of the solution $x(t)$ by $x_*(t)$ of (3.1), we have solved numerically, with high accuracy, the system (3.3) for the FPU problem with the data of Sect. I.5.1. In Figure 3.1 we plot the oscillatory energy $I = I_1 + I_2 + I_3$ with x replaced by the approximation x_* in the definition (2.5) of these quantities. The figure agrees rather well with Figure I.5.2.

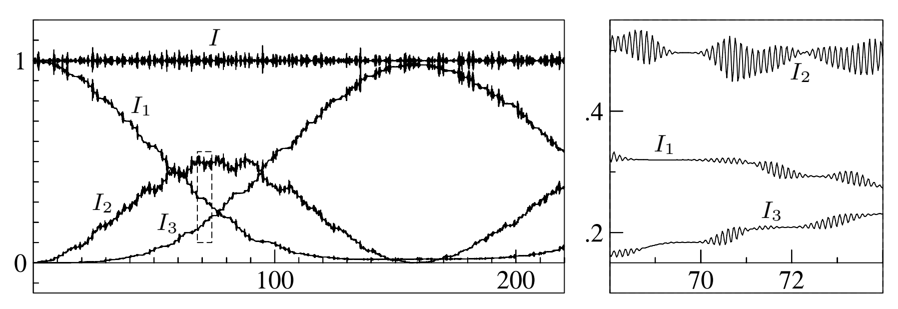


Fig. 3.1. Same experiment as in Fig. I.5.2 for the solution (3.1) of (3.3)

XIII.3.2 Decomposition of the Numerical Solution

For the numerical method (2.6), which solves linear equations $\ddot{x} = -\Omega^2 x$ exactly, we look similarly to the above for a function of the form

$$x_h(t) = y_h(t) + e^{i\omega t} z_h(t) + e^{-i\omega t}\overline{z}_h(t) \tag{3.7}$$

with coefficient functions $y_h(t)$, $z_h(t)$ which are smooth in the sense that all their derivatives are bounded independently of h and ω, such that $x_h(t)$ gives a small defect when inserted into the difference scheme (2.6) and has the correct starting values:

$$x_h(0) = x_0\,, \quad x_h(h) = x_1\,. \tag{3.8}$$

Taylor expansion of $z_h(t \pm h)$ at the point t shows, after some calculation,

$$\begin{aligned}\frac{1}{h^2}\Big(x_h(t+h) - 2\cos(h\Omega)x_h(t) + x_h(t-h)\Big) &= \begin{pmatrix} \delta_h^2 y_{h,0}(t) \\ \sigma_1^2\omega^2 y_{h,1}(t) + \delta_h^2 y_{h,1}(t)\end{pmatrix} \\ &+ e^{i\omega t}\begin{pmatrix} -\sigma_1^2\omega^2 z_{h,0}(t) + \sigma_2\, 2i\omega\dot{z}_{h,0}(t) + \cos(h\omega)\ddot{z}_{h,0}(t) + \ldots \\ \sigma_2\, 2i\omega\dot{z}_{h,1}(t) + \cos(h\omega)\ddot{z}_{h,1}(t) + \ldots\end{pmatrix}\end{aligned} \tag{3.9}$$

$+$ *the complex conjugate of the expression in the previous line,*

where $y_h(t) = \big(y_{h,0}(t), y_{h,1}(t)\big)$ and $z_h(t) = \big(z_{h,0}(t), z_{h,1}(t)\big)$ according to the partitioning in (2.2),

$$\delta_h^2 y_h(t) = \frac{1}{h^2}\Big(y_h(t+h) - 2y_h(t) + y_h(t-h)\Big)$$

is the symmetric second-order difference quotient, $\sigma_k = \operatorname{sinc}(\frac{1}{2}kh\omega)$, and the dots stand for higher powers of h multiplied by derivatives of $z_{h,0}$ or $z_{h,1}$. Taylor expansion of the nonlinearity now gives

$$\begin{aligned}\Psi g(\Phi x_h) &= \Psi g(\Phi y_h) + \Psi g''(\Phi y_h)(\Phi z_h, \Phi\overline{z}_h) \\ &\quad + e^{i\omega t}\Psi g'(\Phi y_h)\Phi z_h + e^{-i\omega t}\Psi g'(\Phi y_h)\Phi\overline{z}_h + \ldots\end{aligned} \tag{3.10}$$

Modified Equations for the Numerical Coefficient Functions. For the moment we consider the case where the absolute values of σ_1 and σ_2 are bounded from below by a positive constant, so that $h\omega$ is assumed bounded and bounded away from a non-zero integral multiple of π. We also assume $h\omega$ to be bounded away from zero, which is the computational situation of interest. In this case, the first term in each line of each bracket in (3.9) can be considered as the dominant one. We therefore require that the functions y_h, z_h satisfy

$$\begin{aligned}\delta_h^2 y_{h,0} &= g_0(\Phi y_h) + g_0''(\Phi y_h)(\Phi z_h, \Phi\overline{z}_h) \\ \operatorname{sinc}^2(\tfrac{1}{2}h\omega)\,\omega^2 y_{h,1} &= \psi(h\omega)\, g_1(\Phi y_h) \\ -\operatorname{sinc}^2(\tfrac{1}{2}h\omega)\,\omega^2 z_{h,0} &= \frac{\partial g_0}{\partial x_0}(\Phi y_h) z_{h,0} + \frac{\partial g_0}{\partial x_1}(\Phi y_h)\phi(h\omega) z_{h,1} \\ \operatorname{sinc}(h\omega)\, 2i\omega\, \dot{z}_{h,1} &= \psi(h\omega)\frac{\partial g_1}{\partial x_0}(\Phi y_h) z_{h,0} + \psi(h\omega)\frac{\partial g_1}{\partial x_1}(\Phi y_h)\phi(h\omega) z_{h,1}\,.\end{aligned} \tag{3.11}$$

The first equation should be stated more precisely as $y_{h,0}$ being a solution of a modified equation for the Störmer–Verlet method (see Exercise IX.3) applied to the corresponding differential equation:

$$\ddot{y}_{h,0} = \left(1 - \frac{h^2}{12}\frac{d^2}{dt^2}\right)\Big(g_0(\Phi y_h) + g_0''(\Phi y_h)(\Phi z_h, \Phi \overline{z}_h)\Big) ,$$

where the time derivatives of $y_{h,1}, z_h$ that result from applying the chain rule are replaced by using the expressions in (3.11). As long as $y_{h,0}(t)$ remains in a bounded domain and $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$, we have again bounds of the same type as for the coefficients of the exact solution:

$$y_{h,1}(t) = \mathcal{O}(\omega^{-2}) , \quad z_{h,0}(t) = \mathcal{O}(\omega^{-3}) , \quad \dot{z}_{h,1}(t) = \mathcal{O}(\omega^{-2}) . \tag{3.12}$$

Initial Values. We next determine the initial values $y_{h,0}(0)$, $\dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ such that $x_h(0)$ and $x_h(h)$ coincide with the starting values $x_0 = x(0)$ and x_1 of the numerical method. We let x_1 be computed from x_0 and $\dot{x}_0$ via the formula (2.7) with $n = 0$, and we still assume that σ_1 and σ_2 are bounded away from zero. Using (3.11), the condition $x_h(0) = x_0 = (x_{0,0}, x_{0,1})$ then becomes

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) + \mathcal{O}\big(\omega^{-2} z_{h,1}(0)\big) \\ x_{0,1} &= z_{h,1}(0) + \overline{z}_{h,1}(0) + \mathcal{O}(\omega^{-2}) . \end{aligned} \tag{3.13}$$

The formula for the first component of (2.7), $x_{1,0} - x_{0,0} = h\dot{x}_{0,0} + \frac{1}{2}h^2 g_0(\Phi x_0)$, together with $x_{h,0}(h) - x_{h,0}(0) = h\dot{y}_{h,0}(0) + \frac{1}{2}h^2 g_0(\Phi x_0) + \mathcal{O}(h^3) + \mathcal{O}(\omega^{-2} z_{h,1}(0))$ implies that

$$\dot{x}_{0,0} = \dot{y}_{h,0}(0) + \mathcal{O}(h^2) + \mathcal{O}\big(\omega^{-1} z_{h,1}(0)\big). \tag{3.14}$$

For the second component we have from (2.7)

$$x_{1,1} - \cos(h\omega) x_{0,1} = h \operatorname{sinc}(h\omega) \dot{x}_{0,1} + \tfrac{1}{2} h^2 \psi(h\omega)\, g_1(\Phi x_0),$$

and by Taylor expansion and (3.11),

$$\begin{aligned} x_{h,1}(h) - \cos(h\omega) x_{h,1}(0) &= (1 - \cos(h\omega))\, y_{h,1}(0) + \mathcal{O}(h\omega^{-2}) \\ &\quad + i \sin(h\omega)\big(z_{h,1}(0) - \overline{z}_{h,1}(0)\big) + \mathcal{O}(h\omega^{-1} z_{h,1}(0)), \end{aligned}$$

where we note the relation $(1 - \cos(h\omega))\, y_{h,1}(0) = \frac{1}{2}h^2 \psi(h\omega)\, g_1(\Phi y_h(0))$ by (3.11) and a trigonometric identity. After division by $h \operatorname{sinc} h\omega = \omega^{-1} \sin h\omega$ the above formulas yield

$$\dot{x}_{0,1} = i\omega\big(z_{h,1}(0) - \overline{z}_{h,1}(0)\big) + \mathcal{O}(\omega^{-2}) + \mathcal{O}\big(\omega^{-1} z_{h,1}(0)\big). \tag{3.15}$$

The four equations (3.13), (3.14), (3.15) constitute a nonlinear system for the four quantities $y_0(0)$, $\dot{y}_0(0)$, $\omega\big(z_{h,1}(0) + \overline{z}_{h,1}(0)\big)$, and $\omega\big(z_{h,1}(0) - \overline{z}_{h,1}(0)\big)$. By fixed-point iteration and using the bounded-energy assumption (2.3), we get a locally unique solution for sufficiently small h, with $z_{h,1}(0) = \mathcal{O}(\omega^{-1})$ and hence

$$
\begin{aligned}
y_{h,0}(0) &= x_{0,0} + \mathcal{O}(\omega^{-3}) , & \dot y_{h,0}(0) &= \dot x_{0,0} + \mathcal{O}(h^2) \\
2\,\mathrm{Re}\, z_{h,1}(0) &= x_{0,1} + \mathcal{O}(\omega^{-2}) , & -\omega\, 2\,\mathrm{Im}\, z_{h,1}(0) &= \dot x_{0,1} + \mathcal{O}(h\omega^{-1}) .
\end{aligned}
\tag{3.16}
$$

Defect. As long as $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$, the defect

$$
d_h(t) = \frac{1}{h^2}\Big(x_h(t+h) - 2\cos(h\Omega) x_h(t) + x_h(t-h)\Big) - \Psi g\big(\Phi x_h(t)\big) \tag{3.17}
$$

is of size $\mathcal{O}(h^2)$ by (3.9)–(3.10) and the very construction (3.11) of the coefficient functions. This estimate refers again to the non-resonant case where σ_1, σ_2 are bounded away from zero and hence $h\omega$ is bounded away from non-zero integral multiples of π. The case of $h\omega$ near a multiple of π requires a special treatment and will be considered in the next subsection.

XIII.4 Accuracy and Slow Exchange

A comparison of the principal terms of the modulated Fourier expansions of the numerical and the exact solution gives much insight into the behaviour of the numerical method and the role of the filter functions ψ and ϕ. From this comparison we obtain error bounds over finite time intervals, and we discuss the slow energy exchange between oscillatory components and the slow energy transfer from oscillatory to smooth components which take place on the time scale ω.

XIII.4.1 Convergence Properties on Bounded Time Intervals

As a first application of the modulated Fourier expansion we consider error bounds on bounded time intervals. Second-order convergence estimates for more general equations $\ddot x = -Ax + g(x)$ with symmetric positive semi-definite matrix A, uniformly in the (arbitrarily large) eigenfrequencies of A, are given by García-Archilla, Sanz-Serna & Skeel (1999) for the mollified impulse method, by Hochbruck & Lubich (1999a) for Gautschi-type methods, and by Grimm & Hochbruck (2005) for general methods of the class (2.7)–(2.8) with appropriate filter functions. Those results were proved with different techniques. The following bounds on the filter functions ψ and ϕ are needed for second-order error bounds of method (2.6):

$$
\begin{aligned}
|\psi(h\omega)| &\le C_1\, \mathrm{sinc}^2(\tfrac12 h\omega) , \\
|\phi(h\omega)| &\le C_2\, |\mathrm{sinc}(\tfrac12 h\omega)| , \\
|\psi(h\omega)\phi(h\omega)| &\le C_3\, |\mathrm{sinc}(h\omega)| .
\end{aligned}
\tag{4.1}
$$

Theorem 4.1. *Consider the numerical solution of the system (2.1) – (2.3) by method (2.6) with a step size $h \le h_0$ (with a sufficiently small h_0 independent of ω) for which $h\omega \ge c_0 > 0$. Let the starting value x_1 be given by (2.7) with $n = 0$. If the conditions (4.1) are satisfied, then the error is bounded by*

$$\|x_n - x(nh)\| \le C\,h^2 \quad \textit{for} \quad nh \le T\,.$$

If only $|\psi(h\omega)| \le C_0\,|\mathrm{sinc}(\frac{1}{2}h\omega)|$ *holds instead of (4.1), then the order of convergence reduces to one:* $\|x_n - x(nh)\| \le C\,h$ *for* $nh \le T$*. In both cases,* C *is independent of* ω*,* h *and* n *with* $nh \le T$ *and of bounds of solution derivatives, but depends on* T*, on* E *of (2.3), on bounds of derivatives of the nonlinearity* g*, and on* C_1, C_2, C_3 *or* C_0*.*

To obtain second-order error bounds uniformly in $h\omega$, condition (4.1) requires a double zero of ψ and a zero of ϕ at even multiples of π, and a zero of ψ or ϕ at odd multiples of π. This is satisfied for the mollified impulse method with $\phi(\xi) = \mathrm{sinc}(\xi)$, for which $\psi(\xi) = \mathrm{sinc}^2(\xi)$. Gautschi-type methods have $\psi(\xi) = \mathrm{sinc}^2(\frac{1}{2}\xi)$, so that the first condition on ψ in (4.1) is trivially satisfied. The conditions on ϕ hold, for example, for $\phi = \mathrm{sinc}$ or for ϕ of (1.20). The original Gautschi method has $\phi = 1$, which does not satisfy the second condition of (4.1), and the Deuflhard/impulse method ($\psi = \mathrm{sinc}, \phi = 1$) satisfies only the third condition of (4.1). These latter methods are not of second order uniformly in $h\omega$.

Proof of Theorem 4.1. (a) First we consider the case where $h\omega$ is bounded away from integral multiples of π, so that condition (4.1) is not needed. Comparing the equations (3.3) and (3.11), which determine the modulated Fourier expansion coefficients, shows

$$y_h(t) - y(t) = \mathcal{O}(h^2)\,, \qquad z_h(t) - z(t) = \mathcal{O}(h^2)$$

on bounded intervals, and hence

$$x_h(t) - x_*(t) = \mathcal{O}(h^2)\,. \tag{4.2}$$

The variation-of-constants formula (1.8) and a Gronwall-type inequality show that, on bounded intervals, the error $x_*(t) - x(t)$ is of the same magnitude as the defect: by (3.6),

$$x_*(t) - x(t) = \mathcal{O}(\omega^{-2})\,.$$

The errors $e_n = x_n - x_h(t_n)$ satisfy

$$e_{n+1} - 2\cos(h\Omega)\,e_n + e_{n-1} = b_n \tag{4.3}$$

with $b_n = h^2\big(\Psi g(\Phi x_n) - \Psi g(\Phi x_h(t_n)) - d_h(t_n)\big)$. This recurrence relation can be solved to yield (Exercise 5)

$$e_{n+1} = -W_{n-1}e_0 + W_n e_1 + \sum_{j=1}^{n} W_{n-j} b_j \tag{4.4}$$

with

$$W_n = \begin{pmatrix} (n+1)I & 0 \\ 0 & \dfrac{\sin(n+1)h\omega}{\sin h\omega}\,I \end{pmatrix}.$$

A discrete Gronwall inequality now yields that on bounded intervals, e_n is of the same magnitude as the defect $d_h(t)$ of (3.17), which is $\mathcal{O}(h^2)$ by the construction of (3.11) and by $z_{h,1} = \mathcal{O}(\omega^{-1})$. Hence,

$$x_n - x_h(t_n) = \mathcal{O}(h^2) \; .$$

Combining these estimates yields the desired second-order error bound.

(b) We now consider the case where $\omega|\mathrm{sinc}(\frac{1}{2}h\omega)| \geq c$ with a sufficiently large constant c, which depends only on bounds of derivatives of g. This condition means that $h\omega$ is outside of an $\mathcal{O}(h)$ neighbourhood of integral multiples of 2π. Under conditions (4.1), the equations (3.11) still give

$$y_{h,1}(t) = \mathcal{O}(\omega^{-2}) \; , \quad z_{h,0}(t) = \mathcal{O}(\omega^{-2}) \; , \quad \dot{z}_{h,1}(t) = \mathcal{O}(\omega^{-2}) \tag{4.5}$$

as long as $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$. Here the first condition of (4.1) gives the bound of $y_{h,1}$, the second one the bound of $z_{h,0}$, and the third one the bound of $\dot{z}_{h,1}$. As in Sect. XIII.3.2, we determine the initial values $y_{h,0}(0)$, $\dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ such that $x_h(0)$ and $x_h(h)$ coincide with the starting values x_0 and x_1 of the numerical method. Using once more (4.1), we obtain a system for the initial values similar to (3.13)–(3.15):

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) + \mathcal{O}\big(\omega^{-1} z_{h,1}(0)\big) \\ x_{0,1} &= z_{h,1}(0) + \overline{z}_{h,1}(0) + \mathcal{O}(\omega^{-2}) \\ \dot{x}_{0,0} &= \dot{y}_{h,0}(0) + \mathcal{O}(h) + \mathcal{O}\big(\omega^{-1} z_{h,1}(0)\big) \\ \dot{x}_{0,1} &= i\omega\big(z_{h,1}(0) - \overline{z}_{h,1}(0)\big) + \mathcal{O}(\omega^{-1}) + \mathcal{O}\big(z_{h,1}(0)\big). \end{aligned} \tag{4.6}$$

With the weaker estimates for $z_{h,0}(t)$ and in (4.6) we still obtain estimates for the initial values of the type (3.16) with at most one factor ω^{-1} or h less in the remainder terms. Condition (2.3) implies again $z_1(0) = \mathcal{O}(\omega^{-1})$, which ensures that (4.5) holds for $0 \leq t \leq T$. The defect is then $d_h(t) = \mathcal{O}(h^2)$, and as in part (a) we get the second-order error bound.

(c) Now let $\omega|\mathrm{sinc}(\frac{1}{2}h\omega)| \leq c$, so that $h\omega$ is $\mathcal{O}(h)$ close to a multiple of 2π. In this case we replace the third equation in (3.11) simply by

$$z_{h,0} = 0.$$

Under condition (4.1) we still obtain the bounds (4.5). The initial values are now chosen to satisfy

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) \\ x_{0,1} &= z_{h,1}(0) + \overline{z}_{h,1}(0) + \omega^{-2} \, \frac{\psi(h\omega)}{\mathrm{sinc}^2(\frac{1}{2}h\omega)} \, g_1(\Phi x_0) \\ \dot{x}_{0,0} &= \dot{y}_{h,0}(0) \\ \dot{x}_{0,1} &= i\omega\big(z_{h,1}(0) - \overline{z}_{h,1}(0)\big). \end{aligned} \tag{4.7}$$

They are then bounded as in (b) and, by the arguments used in the determination of the initial values of Sect. XIII.3.2, yield the estimates $x_h(0) = x_0 + \mathcal{O}(h^3)$

and $x_h(h) = x_1 + \mathcal{O}(h^3)$, and again $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$. Since (4.1) implies $\phi(h\omega)z_{h,1} = \mathcal{O}(\omega^{-2})$ in the present situation of $|\mathrm{sinc}(\frac{1}{2}h\omega)| \le c\,\omega^{-1}$, the defect is still $d_h(t) = \mathcal{O}(h^2)$. The bound (4.2) is also seen to hold. Therefore the second-order error bound remains valid in this case.

(d) If only $|\psi(h\omega)| \le |\mathrm{sinc}(\frac{1}{2}h\omega)|$ holds, then we replace the third equation in (3.11) by $z_{h,0} = 0$. If $\omega|\mathrm{sinc}(\frac{1}{2}h\omega)| \le 1$, we also set $y_{h,1} = 0$. The defect is then only $d_h(t) = \mathcal{O}(h)$, which yields the first-order error bound. □

For the velocity approximation, we obtain the following for the method (2.12) or its equivalent formulations.

Theorem 4.2. *Under the conditions of Theorem 4.1, consider the velocity approximation scheme (2.12) with a function ψ_1 satisfying $\psi_1(0) = 1$ and*

$$|\psi_1(h\omega)| \le C_1' \, |\mathrm{sinc}(\tfrac{1}{2}h\omega)| \, . \tag{4.8}$$

Let the starting values satisfy $\dot x_0 = \dot x(0)$ and $\dot x_1 = \dot x(h) + h\sin(h\Omega)a_1 + \mathcal{O}(h^2)$ with $a_1 = \mathcal{O}(1)$. Then, the error in the velocities is bounded by

$$\|\dot x_n - \dot x(nh)\| \le C\,h \quad \text{for} \quad nh \le T \, ,$$

where C is independent of ω, h and n with $nh \le T$ and of bounds of solution derivatives, but depends on T, on E of (2.3), on bounds of derivatives of the nonlinearity g, and on C_1, C_2, C_3 and C_1'.

Proof. (a) By the variation-of-constants formula (1.8), the exact solution satisfies

$$\begin{aligned} &\dot x(t+h) - 2\cos(h\Omega)\dot x(t) + \dot x(t-h) \\ &\qquad = \int_0^h \cos\big((h-s)\Omega\big)\Big(g\big(x(t+s)\big) - g\big(x(t-s)\big)\Big)\, ds \, . \end{aligned}$$

With the modulated Fourier expansion, we write the exact solution as

$$x(t) \;=\; y(t) + e^{i\omega t} z(t) + e^{-i\omega t}\overline{z}(t) + \mathcal{O}(\omega^{-2})$$

to obtain

$$\begin{aligned} &g\big(x(t+s)\big) - g\big(x(t-s)\big) \\ &\qquad = g'\big(y(t)\big)\Big(2s\,\dot y(t) - 4\sin(\omega s)\,\mathrm{Im}\,\big(e^{i\omega t}z(t)\big) + \mathcal{O}(s^2) + \mathcal{O}(\omega^{-2})\Big). \end{aligned}$$

Using the bounds (3.4), abbreviating $g_{i,j} = \partial g_i/\partial x_j$ and omitting the arguments t and $y(t)$ on the right-hand side, we therefore have

$$\begin{aligned} &\dot x(t+h) - 2\cos(h\Omega)\dot x(t) + \dot x(t-h) \\ &\quad = \begin{pmatrix} h^2\, g_{0,0}\,\dot y_0 - 2h^2\,\mathrm{sinc}^2(\frac{1}{2}h\omega)\,\omega\, g_{0,1}\,\mathrm{Im}\,(e^{i\omega t}z_1) + \mathcal{O}(h^3) \\ h^2\,\mathrm{sinc}^2(\frac{1}{2}h\omega)\, g_{1,0}\dot y_0 - 2h^2\,\mathrm{sinc}(h\omega)\,\omega\, g_{1,1}\,\mathrm{Im}\,(e^{i\omega t}z_1) + \mathcal{O}(h^3) \end{pmatrix}. \end{aligned}$$

We now use the discrete variation-of-constants formula (4.4) and partial summation. For example, the expression

$$\sum_{j=1}^{n} \frac{\sin(n+1-j)h\omega}{\sin h\omega}\, \tfrac{1}{2}h^2 \mathrm{sinc}^2(\tfrac{1}{2}h\omega)\, g_{1,0}\big(y(jh)\big)\, \dot y_0(jh)$$

is seen to be $\mathcal{O}(h)$ uniformly in $h\omega$ and for $nh \le T$ by partial summation, using that the function $g_{1,0}(y(t))\dot y_0(t)$ has a bounded derivative and that

$$\frac{\sin(\frac{1}{2}h\omega)}{\sin(h\omega)} \sum_{j=1}^{k} \sin(jh\omega) = \mathcal{O}(k)\ .$$

In this way we obtain

$$\begin{aligned} \dot x(nh) \;&=\; -W_{n-1}\,\dot x(0) + W_n\,\dot x(h) \\ &\quad + \begin{pmatrix} h\sum_{j=1}^{n}(n+1-j)h\, g_{0,0}(y(jh))\,\dot y_0(jh) \\ 0 \end{pmatrix} + \mathcal{O}(h)\ . \end{aligned} \tag{4.9}$$

(b) For the numerical approximation we proceed similarly. Inserting the modulated Fourier expansion of the numerical solution,

$$x_n = y_h(t) + e^{i\omega t} z_h(t) + e^{-i\omega t}\overline{z}_h(t) + \mathcal{O}(h^2) \qquad \text{for} \quad t = nh \le T\ ,$$

into the numerical scheme, we have with (3.12) or (4.5)

$$\begin{aligned} &\dot x_{n+1} - 2\cos(h\omega)\,\dot x_n + \dot x_{n-1} \\ &= h^2 \begin{pmatrix} g_{0,0}\dot y_{h,0} - 2\,\phi(h\omega)\,\mathrm{sinc}(h\omega)\,\omega\, g_{0,1}\,\mathrm{Im}\,(e^{i\omega t} z_{h,1}) + \mathcal{O}(h) \\ \psi_1(h\omega)\, g_{1,0}\dot y_{h,0} - 2\,(\psi_1\phi)(h\omega)\,\mathrm{sinc}(h\omega)\,\omega\, g_{1,1}\,\mathrm{Im}\,(e^{i\omega t} z_{h,1}) + \mathcal{O}(h) \end{pmatrix} \end{aligned}$$

where the functions $g_{i,j}$ are evaluated at $\Phi y_h(t)$ and the argument $t = nh$ is to be inserted in $\dot y_{h,0}$ and $z_{h,1}$. Under the condition (4.8) on ψ_1, we obtain as in (4.9)

$$\begin{aligned} \dot x_n \;&=\; -W_{n-1}\,\dot x_0 + W_n\,\dot x_1 \\ &\quad + \begin{pmatrix} h\sum_{j=1}^{n}(n+1-j)h\, g_{0,0}(\Phi y_h(jh))\,\dot y_{h,0}(jh) \\ 0 \end{pmatrix} + \mathcal{O}(h)\ . \end{aligned} \tag{4.10}$$

Since we know from the estimates (3.12) and from the proof of Theorem 4.1 that $\Phi y_h(t) = y(t) + \mathcal{O}(h^2)$ and $\dot y_h(t) = \dot y(t) + \mathcal{O}(h^2)$, a comparison of (4.9) and (4.10) gives the result. □

XIII.4.2 Intra-Oscillatory and Oscillatory-Smooth Exchanges

In this subsection we turn to the approximation of slow effects that take place on the time scale ω. Since solutions may depart from each other exponentially, we

cannot expect to obtain small point-wise error bounds on such a time scale. Instead, we take recourse to a kind of formal backward error analysis where we require that the equations determining the modulated Fourier expansion coefficients for the numerical method be small perturbations of those for the exact solution. It may be expected that methods with this property – *ceteribus paris* – show a better long-time behaviour, and this is indeed confirmed by the numerical experiments.

In the Fermi–Pasta–Ulam model, the oscillatory energy of the jth stiff spring is

$$I_j = \tfrac{1}{2}\dot{x}_{1,j}^2 + \tfrac{1}{2}\omega^2 x_{1,j}^2 ,$$

where $x_{1,j}$ is the jth component of the lower block x_1 of x. In terms of the modulated Fourier expansion, this is approximately, up to $\mathcal{O}(\omega^{-1})$,

$$I_j \approx \tfrac{1}{2}\big|i\omega z_{1,j}e^{i\omega t} - i\omega \overline{z}_{1,j}e^{-i\omega t}\big|^2 + \tfrac{1}{2}\omega^2\big|z_{1,j}e^{i\omega t} + \overline{z}_{1,j}e^{-i\omega t}\big|^2 = 2\omega^2|z_{1,j}|^2 .$$

The energy exchange between stiff springs as shown in Fig. 2.1 is thus caused by the slow evolution of z_1 determined by (3.3). This should be modeled correctly by the numerical method.

The term $g_0''(y)(z,\overline{z})$ in the differential equation for y_0 in (3.3) is the dominant term by which the oscillations of the stiff springs exert an influence on the smooth motion. A correct incorporation of this term in the numerical method is desirable.

Upon eliminating y_1 and z_0 in (3.3), the differential equations for y_0 and z_1 become, up to $\mathcal{O}(\omega^{-3})$ perturbations on the right-hand sides,

$$\begin{aligned} \ddot{y}_0 &= g_0(y_0, \omega^{-2}g_1(y_0,0)) + \frac{\partial^2 g_0}{\partial x_1^2}(y_0,0)(z_1,\overline{z}_1) \\ 2i\omega\,\dot{z}_1 &= \frac{\partial g_1}{\partial x_1}(y_0,0)\,z_1 . \end{aligned} \tag{4.11}$$

This is to be compared with the analogous equations for the modulated Fourier expansion of the numerical method, which follow from (3.11):

$$\begin{aligned} \delta_h^2 y_{h,0} &= g_0(y_{h,0}, \gamma\,\omega^{-2}g_1(y_{h,0},0)) + \beta\,\frac{\partial^2 g_0}{\partial x_1^2}(y_{h,0},0)(z_{h,1},\overline{z}_{h,1}) \\ 2i\omega\,\dot{z}_{h,1} &= \alpha\,\frac{\partial g_1}{\partial x_1}(y_{h,0},0)\,z_{h,1} \end{aligned} \tag{4.12}$$

with

$$\alpha = \frac{(\psi\phi)(h\omega)}{\mathrm{sinc}(h\omega)}, \quad \beta = \phi(h\omega)^2, \quad \gamma = \frac{(\psi\phi)(h\omega)}{\mathrm{sinc}^2(\frac{1}{2}h\omega)} . \tag{4.13}$$

The differential equation for $z_{h,1}$ is consistent with that for z_1 only if $\alpha = 1$, i.e.,

$$\psi(h\omega)\,\phi(h\omega) = \mathrm{sinc}(h\omega) . \tag{4.14}$$

Among all the methods (2.6) considered, only the Deuflhard/impulse method ($\psi = \mathrm{sinc}$, $\phi = 1$) satisfies this condition. For this method we indeed observe a qualitatively correct approximation of the energy exchange between stiff springs in

Fig. 2.4, but we have also seen that the energy conservation of this method is very sensitive to near-resonances.

A correct modeling of the slow oscillatory–smooth transfer would in addition require $\beta = 1$ and possibly $\gamma = 1$. For general $h\omega$ the condition $\gamma = 1$ is, however, incompatible with (4.14).

Multi-force methods (2.13) offer a way out of these difficulties. For such methods, the coefficients of the modulated Fourier expansion satisfy (4.12) with (4.13) replaced by

$$\begin{aligned} \alpha \;&=\; \frac{\sum_j \psi_j(h\omega)\,\phi_j(h\omega)}{\mathrm{sinc}(h\omega)}\,, \quad \beta \;=\; \sum_j \psi_j(0)\,\phi_j(h\omega)^2\,, \\ \gamma \;&=\; \sum_j \psi_j(0)\,\phi_j(h\omega)\,\frac{\sum_k \psi_k(h\omega)}{\mathrm{sinc}^2(\frac{1}{2}h\omega)}\,. \end{aligned} \tag{4.15}$$

The two-force method (1.23) with (1.25) has $\alpha = \beta = \gamma = 1$ as desired.

XIII.5 Modulated Fourier Expansions

The decomposition of the exact and the numerical solution into modulated exponentials and a remainder, as derived in Sect. XIII.3, was found useful for understanding several important aspects of the numerical behaviour. Those few terms are, however, not sufficient for explaining the long-time near-conservation of the total and the oscillatory energy. The expansion can be made more accurate by adding further terms $e^{\pm 2i\omega t}$, $e^{\pm 3i\omega t}$ etc. multiplied by slowly varying functions. This leads to an asymptotic expansion which we call the *modulated Fourier expansion*. This expansion is constructed in the present section, following Hairer & Lubich (2000a). (In that paper the modulated Fourier expansion was called the frequency expansion.)

XIII.5.1 Expansion of the Exact Solution

The following theorem extends the construction of Sect. XIII.3.1 to arbitrary order in ω^{-1}.

Theorem 5.1. *Consider a solution $x(t)$ of (2.1) which satisfies the bounded-energy condition (2.3) and stays in a compact set K for $0 \le t \le T$. Then, the solution admits an expansion*

$$x(t) \;=\; y(t) + \sum_{0<|k|<N} e^{ik\omega t} z^k(t) + R_N(t) \tag{5.1}$$

for arbitrary $N \ge 2$, where the remainder term and its derivative are bounded by

$$R_N(t) = \mathcal{O}(\omega^{-N-2}) \quad \textit{and} \quad \dot R_N(t) = \mathcal{O}(\omega^{-N-1}) \qquad \textit{for} \quad 0 \le t \le T\,. \tag{5.2}$$

The real-valued functions $y = (y_0, y_1)$ and the complex-valued functions $z^k = (z_0^k, z_1^k)$ together with all their derivatives (up to arbitrary order M) are bounded by

$$\begin{aligned} y_0 &= \mathcal{O}(1), & z_0^1 &= \mathcal{O}(\omega^{-3}), & z^k &= \mathcal{O}(\omega^{-k-2}) \\ y_1 &= \mathcal{O}(\omega^{-2}), & z_1^1 &= \mathcal{O}(\omega^{-1}), \end{aligned} \tag{5.3}$$

for $k = 2, \ldots, N-1$. Moreover, $z^{-k} = \overline{z^k}$ for all k. These functions are unique up to terms of size $\mathcal{O}(\omega^{-N-2})$. The constants symbolized by the $\mathcal{O}$-notation are independent of ω and t with $0 \le t \le T$ (but they depend on N, T, on E of (2.3), on bounds of the derivatives of the nonlinearity $g(x)$ on K, and on the maximum order M of considered derivatives).

Proof. We set

$$x_*(t) \;=\; y(t) + \sum_{0<|k|<N} e^{ik\omega t} z^k(t) \tag{5.4}$$

and determine the smooth functions $y(t)$, $z(t) = z^1(t)$, and $z^2(t), \ldots, z^{N-1}(t)$ such that $x_*(t)$ inserted into the differential equation (2.1) has a small defect, of size $\mathcal{O}(\omega^{-N})$. To this end we expand $g(x_*(t))$ around $y(t)$ and compare the coefficients of $e^{ik\omega t}$. With the notation $g^{(m)}(y)z^\alpha = g^{(m)}(y)(z^{\alpha_1}, \ldots, z^{\alpha_m})$ for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_m)$, there results the following system of differential equations:

$$\begin{pmatrix} \ddot y_0 \\ \omega^2 y_1 \end{pmatrix} + \begin{pmatrix} 0 \\ \ddot y_1 \end{pmatrix} \;=\; g(y) \;+\; \sum_{s(\alpha)=0} \frac{1}{m!}\, g^{(m)}(y) z^\alpha \tag{5.5}$$

$$\begin{pmatrix} -\omega^2 z_0 \\ 2i\omega \dot z_1 \end{pmatrix} + \begin{pmatrix} 2i\omega \dot z_0 + \ddot z_0 \\ \ddot z_1 \end{pmatrix} \;=\; \sum_{s(\alpha)=1} \frac{1}{m!}\, g^{(m)}(y) z^\alpha \tag{5.6}$$

$$\begin{pmatrix} -k^2\omega^2 z_0^k \\ (1-k^2)\omega^2 z_1^k \end{pmatrix} + \begin{pmatrix} 2ki\omega \dot z_0^k + \ddot z_0^k \\ 2ki\omega \dot z_1^k + \ddot z_1^k \end{pmatrix} \;=\; \sum_{s(\alpha)=k} \frac{1}{m!}\, g^{(m)}(y) z^\alpha \,. \tag{5.7}$$

Here the sums range over all $m \ge 1$ and all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_m)$ with integers α_j satisfying $0 < |\alpha_j| < N$, which have a given sum $s(\alpha) = \sum_{j=1}^m \alpha_j$.

For large ω, the dominating terms in these differential equations are given by the left-most expressions. However, since the central terms involve higher derivatives, we are confronted with singular perturbation problems. We are interested in smooth functions y, z, z^k that satisfy the system up to a defect of size $\mathcal{O}(\omega^{-N})$. In the spirit of Euler's derivation of the Euler-Maclaurin summation formula (see e.g. Hairer & Wanner 1997) we remove the disturbing higher derivatives by using iteratively the differentiated equations (5.5)-(5.7). This leads to a system

$$\begin{aligned} \ddot y_0 &= \mathcal{F}_0(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}), & \dot z_1 &= \omega^{-1}\mathcal{F}_1(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}) \\ z_0 &= \omega^{-2}\mathcal{G}_0(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}), & y_1 &= \omega^{-2}\mathcal{G}_1(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}) \\ z_0^k &= \omega^{-2}\mathcal{G}_0^k(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}), & z_1^k &= \omega^{-2}\mathcal{G}_1^k(\dot y_0, y, z^1, \ldots, z^{N-1}, \omega^{-1}) \end{aligned}$$

where $\mathcal{F}_j, \mathcal{G}_j, \mathcal{G}_j^k$ are formal series in powers of ω^{-1}. Since we get formal algebraic relations for y_1, z_0, z^k, we can further eliminate these variables in the functions $\mathcal{F}_j, \mathcal{G}_j, \mathcal{G}_j^k$. We finally obtain for y_1, z_1, z^k the algebraic relations

$$\begin{aligned} z_0 &= \omega^{-2}\big(G_{00}(y_0,\dot y_0,z_1)+\omega^{-1}G_{01}(y_0,\dot y_0,z_1)+\ldots\big)\\ y_1 &= \omega^{-2}\big(G_{10}(y_0,\dot y_0,z_1)+\omega^{-1}G_{11}(y_0,\dot y_0,z_1)+\ldots\big)\\ z_0^k &= \omega^{-2}\big(G_{00}^k(y_0,\dot y_0,z_1)+\omega^{-1}G_{01}^k(y_0,\dot y_0,z_1)+\ldots\big)\\ z_1^k &= \omega^{-2}\big(G_{10}^k(y_0,\dot y_0,z_1)+\omega^{-1}G_{11}^k(y_0,\dot y_0,z_1)+\ldots\big) \end{aligned} \tag{5.8}$$

and a system of real second-order differential equations for y_0 and complex first-order differential equations for z_1:

$$\begin{aligned} \ddot y_0 &= F_{00}(y_0,\dot y_0,z_1)+\omega^{-1}F_{01}(y_0,\dot y_0,z_1)+\ldots\\ \dot z_1 &= \omega^{-1}\big(F_{10}(y_0,\dot y_0,z_1)+\omega^{-1}F_{11}(y_0,\dot y_0,z_1)+\ldots\big). \end{aligned} \tag{5.9}$$

At this point we can forget the above derivation and take it as a motivation for the ansatz (5.8)-(5.9), which is truncated after the $\mathcal{O}(\omega^{-N})$ terms. We insert this ansatz and its first and second derivatives into (5.5)-(5.7) and compare powers of ω^{-1}. This yields recurrence relations for the functions F_{jl}^k, G_{jl}^k, which in addition show that these functions together with their derivatives are all bounded on compact sets.

We determine initial values for (5.9) such that the function $x_*(t)$ of (5.4) satisfies $x_*(0)=x_0$ and $\dot x_*(0)=\dot x_0$. Because of the special ansatz (5.8)-(5.9), this gives a system which, by fixed-point iteration, yields (locally) unique initial values $y_0(0)$, $\dot y_0(0)$, $z_1(0)$ satisfying (3.5). The assumption (2.3) implies that $z_1(0)=O(\omega^{-1})$. It further follows from the boundedness of F_{1l} that $z_1(t)=O(\omega^{-1})$ for $0\le t\le T$. Going back to (5.7), it is seen that the functions G_{jl}^k contain at least k times the factor z_1. This implies the stated bounds for all other functions.

It remains to estimate the error $R_N(t)=x(t)-x_*(t)$. For this we consider the solution of (5.8)-(5.9) with the above initial values. By construction, these functions satisfy the system (5.5)-(5.7) up to a defect of $\mathcal{O}(\omega^{-N})$. This gives a defect of size $\mathcal{O}(\omega^{-N})$ when the function $x_*(t)$ of (5.4) is inserted into (2.1). On a finite time interval $0\le t\le T$, this implies $R_N(t)=\mathcal{O}(\omega^{-N})$ and $\dot R_N(t)=\mathcal{O}(\omega^{-N})$. To obtain the slightly sharper bounds (5.2), we apply the above proof with N replaced by $N+2$ and use the bounds (5.3) for z^N and z^{N+1}. □

XIII.5.2 Expansion of the Numerical Solution

Does the numerical solution of (2.1) have a modulated Fourier expansion similar to the analytical solution? This may of course be expected, but in Sect. XIII.3.2 we encountered difficulties in constructing the first terms of the expansion in the situation of a numerical resonance where $h\omega$ is close to an integral multiple of π. We therefore confine the discussion to the non-resonant case. We assume that h and ω^{-1} lie in a subregion of the (h,ω^{-1})-plane of small parameters for which there exists a positive constant c such that

$$|\sin(\tfrac{1}{2}kh\omega)| \geq c\sqrt{h} \qquad \text{for } k=1,\ldots,N, \text{ with } N\geq 2. \tag{5.10}$$

This condition implies that $h\omega$ is outside an $\mathcal{O}(\sqrt{h})$ neighbourhood of integral multiples of π. For given h and ω, the condition imposes a restriction on N. In the following, N is a fixed integer such that (5.10) holds. There is the following numerical analogue of Theorem 5.1.

Theorem 5.2. *Consider the numerical solution of the system (2.1) – (2.3) by method (2.6) with step size h. Let the starting value x_1 be given by (2.7) with $n=0$. Assume $h\omega \geq c_0 > 0$, the non-resonance condition (5.10), and the bounds (4.1) for $\psi(h\omega)$ and $\phi(h\omega)$. Then, the numerical solution admits an expansion*

$$x_n = y_h(t) + \sum_{0<|k|<N} e^{ik\omega t} z_h^k(t) + R_{h,N}(t) \tag{5.11}$$

uniformly for $0\leq t=nh\leq T$. The remainder term is of the form

$$R_{h,N}(t) = t^2 h^N \Psi r(t) \quad \text{with} \quad r(t) = \mathcal{O}\big(\phi(h\omega)^N + h^m\big)\,, \tag{5.12}$$

where $m\geq 0$ can be chosen arbitrarily. The coefficient functions together with all their derivatives (up to some arbitrarily fixed order) are bounded by

$$\begin{array}{lll} y_{h,0} = \mathcal{O}(1), & z_{h,0}^1 = \mathcal{O}(\omega^{-2}), & z_{h,0}^k = \mathcal{O}(\omega^{-k}), \\ y_{h,1} = \mathcal{O}(\omega^{-2}), & z_{h,1}^1 = \mathcal{O}(\omega^{-1}), & z_{h,1}^k = \mathcal{O}(\omega^{-k}) \end{array} \tag{5.13}$$

for $k=2,\ldots,N-1$. Moreover, $z_h^{-k} = \overline{z_h^k}$ for all k. The constants symbolized by the $\mathcal{O}$-notation are independent of ω and h with (5.10), but they depend on E, N, m, c, and T.

The *proof* covers the remainder of this subsection. It constructs a function

$$x_h(t) = y_h(t) + \sum_{0<|k|<N} e^{ik\omega t}\, z_h^k(t) \tag{5.14}$$

with smooth coefficient functions $y_h(t)$ and $z_h^k(t)$, which has a small defect when it is inserted into the numerical scheme (2.6). The following functional calculus is convenient for determining the coefficient functions.

Functional Calculus. Let f be an entire complex function bounded by $|f(\zeta)| \leq C\, e^{\gamma|\zeta|}$. Then,

$$f(hD)x(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}\, h^k\, x^{(k)}(t)$$

converges for every function x which is analytic in a disk of radius $r>\gamma h$ around t. If f_1 and f_2 are two such entire functions, then

$$f_1(hD) f_2(hD) x(t) = (f_1 f_2)(hD) x(t)$$

whenever both sides exist. We note $(hD)^k x(t) = h^k x^{(k)}(t)$ for $k = 0, 1, 2, \ldots$ and $\exp(hD)x(t) = x(t+h)$.

We next consider the application of such an operator to functions of the form $e^{i\omega t}z(t)$. By Leibniz' rule of calculus we have $(hD)^k e^{i\omega t} z(t) = e^{i\omega t}(hD + ih\omega)^k z(t)$. After a short calculation this yields

$$f(hD)e^{i\omega t}z(t) \;=\; e^{i\omega t} f(hD + ih\omega) z(t) \tag{5.15}$$

where $f(hD + ih\omega)z(t) = \sum_{k=0}^{\infty} f^{(k)}(ih\omega)/k! \cdot h^k z^{(k)}(t)$.

An N-times continuously differentiable function x is replaced by its Taylor polynomial of degree $N-1$ at t, and $f(hD)x(t)$ is then considered up to $\mathcal{O}(h^N)$.

Modified Equations for the Coefficient Functions. The difference operator of the numerical method becomes in this notation

$$x(t+h) - 2\cos h\Omega\, x(t) + x(t-h) = (e^{hD} - 2\cos h\Omega + e^{-hD})x(t).$$

We factorize this operator as

$$\begin{aligned} \mathcal{L}(hD) \;&:=\; e^{hD} - 2\cos h\Omega + e^{-hD} \;=\; 2\big(\cos(ihD) - \cos h\Omega\big) \\ &=\; 4\,\sin\big(\tfrac12 h\Omega + \tfrac12 ihD\big)\sin\big(\tfrac12 h\Omega - \tfrac12 ihD\big)\;. \end{aligned} \tag{5.16}$$

The function $x_h(t)$ of (5.14) should formally (up to $\mathcal{O}(h^{N+2})$) satisfy the difference scheme

$$\mathcal{L}(hD)x_h(t) \;=\; h^2 \Psi g\big(\Phi x_h(t)\big). \tag{5.17}$$

We insert the ansatz (5.14), expand the right-hand side into a Taylor series around $\Phi y_h(t)$, and compare the coefficients of $e^{ik\omega t}$. This yields the following formal equations for the functions $y_h(t)$ and $z_h^k(t)$:

$$\begin{aligned} \mathcal{L}(hD)y_h \;&=\; h^2\Psi\Big(g(\Phi y_h) + \sum_{s(\alpha)=0} \frac{1}{m!}\, g^{(m)}(\Phi y_h)(\Phi z_h)^\alpha\Big) \\ \mathcal{L}(hD + ikh\omega)z_h^k \;&=\; h^2\Psi \sum_{s(\alpha)=k} \frac{1}{m!}\, g^{(m)}(\Phi y_h)(\Phi z_h)^\alpha. \end{aligned} \tag{5.18}$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_m)$ is a multi-index as in the proof of Theorem 5.1, $s(\alpha) = \sum_{j=1}^m \alpha_j$, and $(\Phi z)^\alpha$ is an abbreviation for the m-tupel $(\Phi z^{\alpha_1}, \ldots, \Phi z^{\alpha_m})$. To get smooth functions $y_h(t)$ and $z_h^k(t)$ which solve (5.18) up to a small defect, we look at the dominating terms in the Taylor expansions of $\mathcal{L}(hD)$ and $\mathcal{L}(hD + ikh\omega)$. With the abbreviations $s_k = \sin(\frac12 kh\omega)$ and $c_k = \cos(\frac12 kh\omega)$ we obtain

$$\begin{aligned} \mathcal{L}(hD) \;&=\; \begin{pmatrix} 0 & 0 \\ 0 & 4s_1^2 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}(ihD)^2 + \ldots \\ \mathcal{L}(hD + ih\omega) \;&=\; \begin{pmatrix} -4s_1^2 & 0 \\ 0 & 0 \end{pmatrix} + 2s_2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}(ihD) \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad - c_2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD)^2 + \dots \qquad (5.19)\\
\mathcal{L}(hD + ikh\omega) &= \begin{pmatrix} -4s_k^2 & 0 \\ 0 & -4s_{k-1}s_{k+1} \end{pmatrix} + 2s_{2k} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD) \\
&\qquad - c_{2k} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD)^2 + \dots .
\end{aligned}
$$

Construction of the Coefficient Functions. Under the non-resonance condition (5.10), the first non-vanishing coefficients in (5.19) are the dominant ones, and the derivation of the defining relations for y_h and z_h^k is the same as for the analytical solution in Theorem 5.1; see also part (b) of the proof of Theorem 4.1. We insert (5.19) into (5.18) and we eliminate recursively the higher derivatives. This motivates the following ansatz for the computation of the functions y_h and z_h^k:

$$
\begin{aligned}
\ddot{y}_{h,0} &= f_{00}(\cdot) + \sqrt{h}\, f_{01}(\cdot) + h\, f_{02}(\cdot) + \dots \\
\dot{z}_{h,1}^1 &= \frac{\psi(h\omega)h}{s_2} \Big(f_{10}(\cdot) + \sqrt{h}\, f_{11}(\cdot) + \dots \Big) \\
z_{h,0}^1 &= \frac{h^2}{s_1^2} \Big(g_{00}^1(\cdot) + \sqrt{h}\, g_{01}^1(\cdot) + \dots \Big) \\
y_{h,1} &= \frac{\psi(h\omega)h^2}{s_1^2} \Big(g_{10}(\cdot) + \sqrt{h}\, g_{11}(\cdot) + \dots \Big) \qquad (5.20) \\
z_{h,0}^k &= \frac{h^2}{s_k^2} \Big(g_{00}^k(\cdot) + \sqrt{h}\, g_{01}^k(\cdot) + \dots \Big) \\
z_{h,1}^k &= \frac{\psi(h\omega)h^2}{s_{k+1}s_{k-1}} \Big(g_{10}^k(\cdot) + \sqrt{h}\, g_{11}^k(\cdot) + \dots \Big),
\end{aligned}
$$

for $k = 2, \dots, N-1$, where the functions depend smoothly on the variables $y_{h,0}$, $\dot{y}_{h,0}$, $\phi(h\omega)z_{h,1}^1$ and on the bounded parameters $\sqrt{h}/s_k$, s_k, c_k, $\psi(h\omega)$ and $(h\omega)^{-1}$. Inserting this ansatz and its derivatives into (5.18) and comparing powers of $\sqrt{h}$ yields recurrence relations for the functions f_{jl}^k, g_{jl}^k. The functions g_{jl}^k (for $k \geq 1$) contain at least k times the factor $\phi(h\omega)z_{h,1}^1$, and f_{1l} contains this factor at least once. Since the series in (5.20) need not converge, we truncate them after the h^{N+m+2} terms.

Initial Values. The conditions $x_h(0) = x_0$ and $x_h(h) = x_1$ determine the initial values $y_{h,0}(0)$, $\dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ in the same way as in Sect. XIII.3.2. Condition (4.1) yields again (4.6), and (2.3) then implies $z_{h,1}(0) = \mathcal{O}(\omega^{-1})$.

Defect. It follows from (4.1) that $h\psi(h\omega)\phi(h\omega)/s_2 = \mathcal{O}(\omega^{-1})$, so that $\dot{z}_{h,1}^1 = \mathcal{O}(\omega^{-1}z_{h,1}^1)$ by (5.20). This implies $z_{h,1}^1(t) = \mathcal{O}(\omega^{-1})$ for $t \leq T$. The other estimates (5.13) are directly obtained from (5.20), which indeed yields the following more refined bounds for the coefficient functions together with their derivatives:

$$\begin{aligned} &y_{h,0} = \mathcal{O}(1), \quad y_{h,1} = \mathcal{O}(\omega^{-2}) \\ &z_{h,0}^1 = \mathcal{O}(\omega^{-3}/\sqrt{h}), \quad z_{h,1}^1 = \mathcal{O}(\omega^{-1}), \quad \dot{z}_{h,1}^1 = \mathcal{O}(\omega^{-2}) \\ &z_{h,0}^k = \mathcal{O}(h\phi(h\omega)^k\omega^{-k}), \quad z_{h,1}^k = \mathcal{O}(h\psi(h\omega)\phi(h\omega)^k\omega^{-k}). \end{aligned} \tag{5.21}$$

Consequently, the values $x_h(nh)$ inserted into the numerical scheme (2.6) yield a defect of size $\mathcal{O}(h^{N+2})$:

$$\begin{aligned} &x_h(t+h) - 2\cos(h\Omega)\, x_h(t) + x_h(t-h) = \\ &\quad = h^2\, \Psi\Big(g(\Phi x_h(t)) + \mathcal{O}\big(\phi(h\omega)^N\omega^{-N} + h^{N+m}\big)\Big). \end{aligned} \tag{5.22}$$

Standard convergence estimates then show that, on bounded time intervals, $x_n - x_h(nh)$ is of size $\mathcal{O}(t^2h^N)$ and actually satisfies the finer estimate (5.12). This completes the proof of Theorem 5.2. □

XIII.5.3 Expansion of the Velocity Approximation

A similar expansion holds also for the velocities. We show this for the scheme (2.11) or its equivalent one-step formulation (2.8) with (2.9).

Theorem 5.3. *Under the assumptions of Theorem 5.2, the velocity approximation $\dot{x}_n$ given by (2.11) has an expansion*

$$\dot{x}_n \;=\; v_h(t) + \sum_{0<|k|<N} e^{ik\omega t} w_h^k(t) + \mathcal{O}(t^2h^{N-1})$$

uniformly for $0 \le t = nh \le T$, where the real-valued functions $v_h = (v_{h,0}, v_{h,1})$ and the complex-valued functions $w_h^k = (w_{h,0}^k, w_{h,1}^k)$ together with their derivatives up to arbitrary order satisfy

$$\begin{aligned} &v_{h,0} = \dot{y}_{h,0} + \mathcal{O}(h^2), \qquad w_{h,0}^1 = \mathcal{O}(\omega^{-1}), \quad w_{h,0}^k = \mathcal{O}(\omega^{-k}) \\ &w_{h,1}^1 = i\omega z_{h,1}^1 + \mathcal{O}(\omega^{-1}), \quad v_{h,1} = \mathcal{O}(\omega^{-1}), \quad w_{h,1}^k = \mathcal{O}(\omega^{-k}) \end{aligned} \tag{5.23}$$

for $k = 2, \dots, N-1$. Moreover, $w_h^{-k} = \overline{w_h^k}$. The constants symbolized by the $\mathcal{O}$-notation are independent of ω and h with (5.10), but depend on E, N, c, and T.

Proof. Let $u_h(t)$ be defined by the continuous analogue of (2.11),

$$2h\,\mathrm{sinc}(h\Omega)\, u_h(t) \;=\; x_h(t+h) - x_h(t-h). \tag{5.24}$$

Theorem 5.2 then yields that

$$\dot{x}_n \;=\; u_h(t) + \mathcal{O}(t^2h^{N-1})$$

for $t = nh$ on bounded time intervals. Here we used that the remainder term in the lower component of (5.12) is of the form $\mathcal{O}(\psi(h\omega)(\phi(h\omega) + h)t^2h^N)$, so that its

quotient with $2h\,\mathrm{sinc}(h\omega)$ becomes $\mathcal{O}(t^2h^{N-1})$ by the third of the conditions (4.1) and by (5.10). The function $u_h(t)$ can be written as

$$u_h(t) \;=\; v_h(t) + \sum_{0<|k|<N} e^{ik\omega t} w_h^k(t)\;. \tag{5.25}$$

We insert the relation (5.14) into $-i\sin(ihD)x_h(t) \;=\; h\,\mathrm{sinc}(h\Omega)u_h(t)$, which is equivalent to (5.24), and compare the coefficients of $e^{ik\omega t}$ to obtain

$$\begin{aligned}
\mathrm{sinc}(ihD)\,\dot y_{h,0} \;&=\; v_{h,0}\\
\mathrm{sinc}(ihD)\,\dot y_{h,1} \;&=\; \mathrm{sinc}(h\omega)v_{h,1}\\
(ih)^{-1}\,\sin(ihD-kh\omega)\,z_{h,0}^k \;&=\; w_{h,0}^k\\
(ih)^{-1}\,\sin(ihD-kh\omega)\,z_{h,1}^k \;&=\; \mathrm{sinc}(h\omega)w_{h,1}^k
\end{aligned} \tag{5.26}$$

for $k=1,\ldots,N-1$. In particular, for $w_{h,1}^1$ we get

$$w_{h,1}^1 \;=\; i\omega\cos(ihD)\,z_{h,1}^1 - i\omega\,\frac{\cos(h\omega)}{\sin(h\omega)}\,\sin(ihD)\,z_{h,1}^1. \tag{5.27}$$

With the above equations, the estimates now follow with the bounds (5.21) of the coefficient functions and their derivatives, using again (4.1). □

XIII.6 Almost-Invariants of the Modulated Fourier Expansions

The system for the coefficients of the modulated Fourier expansion of the exact solution is shown to have two formal invariants, which are related to the total and the oscillatory energy. In particular, this explains the near-conservation of the oscillatory energy over very long times. Analogous almost-invariants are shown to exist also for the modulated Fourier expansion of the numerical solution. This forms the basis for results on the long-time energy conservation of numerical methods, which will be given in Sections XIII.7 and XIII.8.

XIII.6.1 The Hamiltonian of the Modulated Fourier Expansion

The equation (2.1) is a Hamiltonian system with the Hamiltonian

$$H(x,\dot x) \;=\; \tfrac12 \dot x^T\dot x + \tfrac12 x^T\Omega^2 x + U(x)\;. \tag{6.1}$$

In the modulated Fourier expansion of the solution $x(t)$ of (2.1), denote $y^0(t) = y(t)$ and $y^k(t) = e^{ik\omega t}z^k(t)$ ($0<|k|<N$), and let

$$\mathbf{y} = (y^{-N+1},\ldots,y^{-1},y^0,y^1,\ldots,y^{N-1})\;.$$

By (5.5)–(5.7) these functions satisfy

$$\ddot{y}^k + \Omega^2 y^k \;=\; -\sum_{s(\alpha)=k} \frac{1}{m!}\, U^{(m+1)}(y^0)\, \mathbf{y}^\alpha + \mathcal{O}(\omega^{-N})\,. \tag{6.2}$$

Here, the sum is over all $m \geq 0$ and all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_m)$ with integers α_j $(0 < |\alpha_j| < N)$ which have a given sum $s(\alpha) = \sum_{j=1}^m \alpha_j$, and we write $\mathbf{y}^\alpha = (y^{\alpha_1}, \ldots, y^{\alpha_m})$. We define

$$\mathcal{U}(\mathbf{y}) \;=\; U(y^0) + \sum_{s(\alpha)=0} \frac{1}{m!}\, U^{(m)}(y^0)\, \mathbf{y}^\alpha\,. \tag{6.3}$$

From the above it follows that $\mathbf{y}(t)$ satisfies the system

$$\ddot{y}^k + \Omega^2 y^k \;=\; -\nabla_{y^{-k}}\, \mathcal{U}(\mathbf{y}) + \mathcal{O}(\omega^{-N}) \tag{6.4}$$

which, neglecting the $\mathcal{O}(\omega^{-N})$ term, is the Hamiltonian system (cf. Exercise 6)

$$\dot{y}^k \;=\; \frac{\partial \mathcal{H}}{\partial \dot{y}^{-k}}(\mathbf{y}, \dot{\mathbf{y}}), \qquad \ddot{y}^k \;=\; -\frac{\partial \mathcal{H}}{\partial y^{-k}}(\mathbf{y}, \dot{\mathbf{y}}) \tag{6.5}$$

with

$$\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}}) = \frac{1}{2} \sum_{|k|<N} \Big((\dot{y}^{-k})^T \dot{y}^k + (y^{-k})^T \Omega^2 y^k \Big) + \mathcal{U}(\mathbf{y})\,. \tag{6.6}$$

Theorem 6.1. *Under the assumptions of Theorem 5.1, the Hamiltonian of the modulated Fourier expansion satisfies*

$$\mathcal{H}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) \;=\; \mathcal{H}(\mathbf{y}(0), \dot{\mathbf{y}}(0)) + \mathcal{O}(\omega^{-N}) \tag{6.7}$$

$$\mathcal{H}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) \;=\; H(x(t), \dot{x}(t)) + \mathcal{O}(\omega^{-1})\,. \tag{6.8}$$

The constants symbolized by $\mathcal{O}$ are independent of ω and t with $0 \leq t \leq T$, but depend on E, N and T.

Proof. Multiplying (6.4) with $(\dot{y}^{-k})^T$ and summing up gives

$$\sum_{|k|<N} (\dot{y}^{-k})^T (\ddot{y}^k + \Omega^2 y^k) \;=\; -\frac{d}{dt}\, \mathcal{U}(\mathbf{y}) + \mathcal{O}(\omega^{-N})\,. \tag{6.9}$$

Integrating from 0 to t and using $y^{-k} = \overline{y^k}$ then yields (6.7).

By the bounds of Theorem 5.1, we have for $0 \leq t \leq T$

$$\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}}) = \tfrac{1}{2} \|\dot{y}_0^0\|^2 + \|\dot{y}_1^1\|^2 + \omega^2 \|y_1^1\|^2 + U(y^0) + \mathcal{O}(\omega^{-1}). \tag{6.10}$$

On the other hand, we have from (6.1) and (5.1)

$$H(x, \dot{x}) = \tfrac{1}{2} \|\dot{y}_0^0\|^2 + \tfrac{1}{2} \|\dot{y}_1^1 + \dot{y}_1^{-1}\|^2 + \tfrac{1}{2} \omega^2 \|y_1^1 + y_1^{-1}\|^2 + U(y^0) + \mathcal{O}(\omega^{-1}). \tag{6.11}$$

Using $y_1^1 = e^{i\omega t} z_1^1$ and $\dot{y}_1^1 = e^{i\omega t}(\dot{z}_1^1 + i\omega z_1^1)$ together with $y_1^{-1} = \overline{y_1^1}$, it follows from $\dot{z}_1^1 = \mathcal{O}(\omega^{-1})$ that $\dot{y}_1^1 + \dot{y}_1^{-1} = i\omega(y_1^1 - y_1^{-1}) + \mathcal{O}(\omega^{-1})$ and $\|\dot{y}_1^1\| = \omega \|y_1^1\| + \mathcal{O}(\omega^{-1})$. Inserted into (6.10) and (6.11), this yields (6.8). $\square$

XIII.6.2 A Formal Invariant Close to the Oscillatory Energy

In addition to the Hamiltonian $\mathcal{H}(\mathbf{y},\dot{\mathbf{y}})$, the system for the coefficients of the modulated Fourier expansion has another formally conserved quantity. This almost-invariant depends only on the oscillating part and is given by

$$\mathcal{I}(\mathbf{y},\dot{\mathbf{y}}) \;=\; -\,i\omega \sum_{0<|k|<N} k\,(y^{-k})^T \dot{y}^k. \tag{6.12}$$

This turns out to be close to the energy of the harmonic oscillator,

$$I(x,\dot{x}) \;=\; \tfrac{1}{2}\|\dot{x}_1\|^2 + \tfrac{1}{2}\omega^2\|x_1\|^2\;. \tag{6.13}$$

Theorem 6.2. *Under the assumptions of Theorem 5.1,*

$$\mathcal{I}(\mathbf{y}(t),\dot{\mathbf{y}}(t)) \;=\; \mathcal{I}(\mathbf{y}(0),\dot{\mathbf{y}}(0)) + \mathcal{O}(\omega^{-N}) \tag{6.14}$$

$$\mathcal{I}(\mathbf{y}(t),\dot{\mathbf{y}}(t)) \;=\; I(x(t),\dot{x}(t)) + \mathcal{O}(\omega^{-1})\;. \tag{6.15}$$

The constants symbolized by $\mathcal{O}$ are independent of ω and t with $0 \le t \le T$, but depend on E, N and T.

Proof. For the vector

$$\mathbf{y}(\lambda) = \big(e^{i(-N+1)\lambda}y^{-N+1},\dots,e^{-i\lambda}y^{-1},y^0,e^{i\lambda}y^1,\dots,e^{i(N-1)\lambda}y^{N-1}\big)$$

the definition (6.3) of $\mathcal{U}$ shows that $\mathcal{U}(\mathbf{y}(\lambda))$ does not depend on λ. Its derivative with respect to λ thus yields

$$0 \;=\; \frac{d}{d\lambda}\mathcal{U}(\mathbf{y}(\lambda)) \;=\; \sum_{0<|k|<N} ik\,e^{ik\lambda}(y^k)^T\,\nabla_k\,\mathcal{U}(\mathbf{y}(\lambda))\;,$$

and putting $\lambda=0$ we obtain

$$\sum_{0<|k|<N} ik\,(y^k)^T\,\nabla_k\,\mathcal{U}(\mathbf{y}) \;=\; 0 \tag{6.16}$$

for all vectors $\mathbf{y}=(y^{-N+1},\dots,y^{-1},y^0,y^1,\dots,y^{N-1})$.

The proof of Theorem 6.2 is now very similar to that of Theorem 6.1. We multiply the relation (6.4) with $-i\omega k(y^{-k})^T$ instead of $(\dot{y}^{-k})^T$. Summing up yields, with the use of (6.16),

$$-i\omega \sum_{0<|k|<N} k\,(y^{-k})^T\big(\ddot{y}^k + \Omega^2 y^k\big) \;=\; \mathcal{O}(\omega^{-N})\;. \tag{6.17}$$

The time derivative of $\mathcal{I}(\mathbf{y},\dot{\mathbf{y}})$ of (6.12) equals

$$\frac{d}{dt}\mathcal{I}(\mathbf{y},\dot{\mathbf{y}}) \;=\; -\,i\omega \sum_{0<|k|<N} k\Big((y^{-k})^T\ddot{y}^k + (\dot{y}^{-k})^T\dot{y}^k\Big). \tag{6.18}$$

In the sums $\sum_k k(y^{-k})^T \Omega^2 y^k$ and $\sum_k k(\dot{y}^{-k})^T \dot{y}^k$, the terms with k and $-k$ cancel. Hence, (6.17) and (6.18) together yield

$$\frac{d}{dt}\mathcal{I}(\mathbf{y},\dot{\mathbf{y}}) \;=\; \mathcal{O}(\omega^{-N}) \;,$$

which implies (6.14).

With $\dot{y}^k = e^{ik\omega t}(\dot{z}^k + ik\omega z^k) = ik\omega y^k + \mathcal{O}(\omega^{-1})$, it follows from the bounds of Theorem 5.1 that

$$\mathcal{I}(\mathbf{y},\dot{\mathbf{y}}) \;=\; 2\omega^2 \|y_1^1\|^2 + \mathcal{O}(\omega^{-1}).$$

On the other hand, using the arguments of the proof of Theorem 6.1, we have

$$I(x,\dot{x}) = \tfrac{1}{2}\|\dot{y}_1^1 + \dot{y}_1^{-1}\|^2 + \tfrac{1}{2}\omega^2\|y_1^1 + y_1^{-1}\|^2 + \mathcal{O}(\omega^{-1}) = 2\omega^2\|y_1^1\|^2 + \mathcal{O}(\omega^{-1}).$$

This proves the second statement of the theorem. □

Theorem 6.2 implies that the oscillatory energy is nearly conserved over long times:

Theorem 6.3. *If the solution $x(t)$ of (2.1) stays in a compact set for $0 \le t \le \omega^N$, then*

$$I(x(t),\dot{x}(t)) \;=\; I(x(0),\dot{x}(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(t\omega^{-N}) \;.$$

The constants symbolized by $\mathcal{O}$ are independent of ω and t with $0 \le t \le \omega^N$, but depend on E and N.

Proof. With a fixed $T > 0$, let $\mathbf{y}_j$ denote the vector of the modulated Fourier expansion terms that correspond to starting values $(x(jT),\dot{x}(jT))$. For $t = (n+\theta)T$ with $0 \le \theta < 1$, we have by (6.15)

$$\begin{aligned}
&I(x(t),\dot{x}(t)) - I(x(0),\dot{x}(0)) \\
&= \mathcal{I}(\mathbf{y}_n(\theta T),\dot{\mathbf{y}}_n(\theta T)) + \mathcal{O}(\omega^{-1}) - \mathcal{I}(\mathbf{y}_0(0),\dot{\mathbf{y}}_0(0)) + \mathcal{O}(\omega^{-1}) \\
&= \mathcal{I}(\mathbf{y}_n(\theta T),\dot{\mathbf{y}}_n(\theta T)) - \mathcal{I}(\mathbf{y}_n(0),\dot{\mathbf{y}}_n(0)) + \\
&\quad \sum_{j=0}^{n-1}\Big(\mathcal{I}(\mathbf{y}_{j+1}(0),\dot{\mathbf{y}}_{j+1}(0)) - \mathcal{I}(\mathbf{y}_j(0),\dot{\mathbf{y}}_j(0))\Big) + \mathcal{O}(\omega^{-1}) \;.
\end{aligned}$$

We note

$$\mathcal{I}(\mathbf{y}_{j+1}(0),\dot{\mathbf{y}}_{j+1}(0)) - \mathcal{I}(\mathbf{y}_j(0),\dot{\mathbf{y}}_j(0)) \;=\; \mathcal{O}(\omega^{-N}) \;,$$

because, by the quasi-uniqueness of the coefficient functions as stated by Theorem 5.1, we have $\mathbf{y}_{j+1}(0) = \mathbf{y}_j(T) + \mathcal{O}(\omega^{-N})$ and $\dot{\mathbf{y}}_{j+1}(0) = \dot{\mathbf{y}}_j(T) + \mathcal{O}(\omega^{-N})$, and we have the bound (6.14) of Theorem 6.2. The same argument applies to $\mathcal{I}(\mathbf{y}_n(\theta T),\dot{\mathbf{y}}_n(\theta T)) - \mathcal{I}(\mathbf{y}_n(0),\dot{\mathbf{y}}_n(0))$. This yields the result. □

In a different approach, Benettin, Galgani & Giorgilli (1987) use a sequence of coordinate transformations from Hamiltonian perturbation theory to show that I has only small deviations over time intervals which grow exponentially with ω, in the case of an analytic potential U. By carefully tracing the dependence on N of the constants in the $\mathcal{O}(\omega^{-N})$-terms, near-conservation of I over exponentially long time intervals can be shown also within the present framework of modulated Fourier expansions; see Cohen, Hairer & Lubich (2003).

XIII.6.3 Almost-Invariants of the Numerical Method

We show that the coefficients of the modulated Fourier expansion of the numerical solution have almost-invariants that are obtained similarly to the above. We denote

$$\mathbf{y}_h = (y_h^{-N+1}, \dots, y_h^{-1}, y_h^0, y_h^1, \dots, y_h^{N-1})$$
$$\mathbf{z}_h = (z_h^{-N+1}, \dots, z_h^{-1}, z_h^0, z_h^1, \dots, z_h^{N-1})$$

with $y_h^0(t) = z_h^0(t) = y_h(t)$ and $y_h^k(t) = e^{ik\omega t} z_h^k(t)$, where y_h and z_h^k are the coefficients of the modulated Fourier expansion of Theorem 5.2. Similar to (6.3) we consider the function

$$\mathcal{U}_h(\mathbf{y}_h) \;=\; U(\Phi y_h^0) + \sum_{s(\alpha)=0} \frac{1}{m!}\, U^{(m)}(\Phi y_h^0)(\Phi \mathbf{y}_h)^\alpha \,, \tag{6.19}$$

where the sum is again taken over all $m \geq 1$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with $0 < |\alpha_j| < N$ for which $s(\alpha) = \sum_j \alpha_j = 0$. It then follows from (5.22), multiplied with $h^{-2}\Psi^{-1}\Phi$, that the functions $y_h^k(t)$ satisfy

$$\Psi^{-1}\Phi\, h^{-2}\mathcal{L}(hD)\, y_h^k \;=\; -\nabla_{-k}\, \mathcal{U}_h(\mathbf{y}_h) + \mathcal{O}(h^N)\,, \tag{6.20}$$

where $\mathcal{L}(hD)$ of (5.16) denotes again the difference operator of the numerical method. The similarity of these relations to (6.4) allows us to obtain almost-conserved quantities that are analogues of $\mathcal{H}$ and $\mathcal{I}$ above.

The First Almost-Invariant. We multiply (6.20) by $(\dot y_h^{-k})^T$, and as in (6.9) we obtain

$$\sum_{|k|<N} (\dot y_h^{-k})^T \Psi^{-1}\Phi\, h^{-2}\mathcal{L}(hD) y_h^k + \frac{d}{dt}\mathcal{U}_h(\mathbf{y}_h) \;=\; \mathcal{O}(h^N)\,.$$

Since we know bounds of the coefficient functions z_h^k and of their derivatives from Theorem 5.2, we switch to the quantities z_h^k and we get the equivalent relation

$$\sum_{|k|<N} (\dot z_h^{-k} - ik\omega z_h^{-k})^T \Psi^{-1}\Phi h^{-2}\mathcal{L}(hD + ik\omega h) z_h^k + \frac{d}{dt}\,\mathcal{U}_h(\mathbf{z}_h) = \mathcal{O}(h^N). \tag{6.21}$$

We shall show that the left-hand side is the total derivative of an expression that depends only on z_h^k and derivatives thereof. Consider first the term for $k = 0$. The

symmetry of the numerical method enters at this very point in the way that the expression $\mathcal{L}(hD)y = h^2\ddot{y} + c_4h^4y^{(4)} + c_6h^6y^{(6)} + \ldots$ contains only terms with derivatives of an even order. Multiplied with $\dot{y}^T$, even-order derivatives of y give a total derivative:

$$\dot{y}^T y^{(2l)} \;=\; \frac{d}{dt}\Big(\dot{y}^T y^{(2l-1)} - \ddot{y}^T y^{(2l-2)} + \ldots \mp (y^{(l-1)})^T y^{(l+1)} \pm \frac{1}{2}\,(y^{(l)})^T y^{(l)}\Big).$$

Thanks to the symmetry of the difference operator $\mathcal{L}(hD)$ only expressions of this type appear in the term for $k=0$ in (6.21), with z_h^0 in the role of y. Similarly, we get for $z = z_h^k$ and $\overline{z} = z_h^{-k}$ with $0 < |k| < N$

$$\begin{aligned}
\operatorname{Re}\dot{\overline{z}}^T z^{(2l)} &= \operatorname{Re}\frac{d}{dt}\Big(\dot{\overline{z}}^T z^{(2l-1)} - \ldots \mp (\overline{z}^{(l-1)})^T z^{(l+1)} \pm \frac{1}{2}\,(\overline{z}^{(l)})^T z^{(l)}\Big)\\
\operatorname{Re}\overline{z}^T z^{(2l+1)} &= \operatorname{Re}\frac{d}{dt}\Big(\overline{z}^T z^{(2l)} - \ldots \pm (\overline{z}^{(l-1)})^T z^{(l+1)} \mp \frac{1}{2}\,(\overline{z}^{(l)})^T z^{(l)}\Big)\\
\operatorname{Im}\dot{\overline{z}}^T z^{(2l+1)} &= \operatorname{Im}\frac{d}{dt}\Big(\dot{\overline{z}}^T z^{(2l)} - \ddot{\overline{z}}^T z^{(2l-1)} + \ldots \mp (\overline{z}^{(l)})^T z^{(l+1)}\Big)\\
\operatorname{Im}\overline{z}^T z^{(2l+2)} &= \operatorname{Im}\frac{d}{dt}\Big(\overline{z}^T z^{(2l+1)} - \dot{\overline{z}}^T z^{(2l)} + \ldots \pm (\overline{z}^{(l)})^T z^{(l+1)}\Big)\,.
\end{aligned}$$

Using the formulas (5.19) for $\mathcal{L}(hD + ikh\omega)$, it is seen that the term for k in (6.21) has an asymptotic h-expansion with expressions of the above type as coefficients. The left-hand side of (6.21) can therefore be written as the time derivative of a function $\widehat{\mathcal{H}}_h[\mathbf{z}_h](t)$ which depends on the values at t of the coefficient function vector $\mathbf{z}_h$ and its first N time derivatives. The relation (6.21) thus becomes

$$\frac{d}{dt}\widehat{\mathcal{H}}_h[\mathbf{z}_h](t) \;=\; \mathcal{O}(h^N)\,.$$

Together with the estimates of Theorem 5.2, this construction of $\widehat{\mathcal{H}}_h$ yields the following result.

Lemma 6.4. *Under the assumptions of Theorem 5.2, the coefficient functions* $\mathbf{z}_h = (z_h^{-N+1}, \ldots, z_h^{-1}, y_h, z_h^1, \ldots, z_h^{N-1})$ *of the modulated Fourier expansion of the numerical solution satisfy*

$$\widehat{\mathcal{H}}_h[\mathbf{z}_h](t) \;=\; \widehat{\mathcal{H}}_h[\mathbf{z}_h](0) \;+\; \mathcal{O}(th^N) \tag{6.22}$$

for $0 \le t \le T$. *Moreover,*

$$\widehat{\mathcal{H}}_h[\mathbf{z}_h](t) \;=\; \tfrac{1}{2}\,\|\dot{y}_{h,0}(t)\|^2 + \sigma(h\omega)\,2\omega^2\,\|z_{h,1}^1(t)\|^2 + U(\Phi y_h(t)) + \mathcal{O}(h^2), \tag{6.23}$$

where $\sigma(h\omega) = \operatorname{sinc}(h\omega)\phi(h\omega)/\psi(h\omega)$. □

The Second Almost-Invariant. By the same calculation as in the proof of Theorem 6.2 we obtain for $\mathcal{U}_h(\mathbf{y}_h(t))$ of (6.19)

$$0 = \sum_{0<|k|<N} ik\omega\, (y_h^k)^T\, \nabla_k\, \mathcal{U}_h(\mathbf{y}_h)\,.$$

It then follows from (6.20) that

$$-\, i\omega \sum_{0<|k|<N} k(y_h^{-k})^T \Psi^{-1} \Phi\, h^{-2} \mathcal{L}(hD) y_h^k \;=\; \mathcal{O}(h^N)\,.$$

Written in the z variables, this becomes

$$-\, i\omega \sum_{0<|k|<N} k(z_h^{-k})^T \Psi^{-1} \Phi\, h^{-2} \mathcal{L}(hD + ik\omega h) z_h^k \;=\; \mathcal{O}(h^N)\,. \tag{6.24}$$

As in (6.21), the left-hand expression can be written as the time derivative of a function $\widehat{\mathcal{I}}_h[\mathbf{z}_h](t)$ which depends on the values at t of the function $\mathbf{z}_h$ and its first N derivatives:

$$\frac{d}{dt}\widehat{\mathcal{I}}_h[\mathbf{z}_h](t) \;=\; \mathcal{O}(h^N)\,.$$

Together with the estimates of Theorem 5.2 this yields the following result.

Lemma 6.5. *Under the assumptions of Theorem 5.2, the coefficient functions* $\mathbf{z}_h$ *of the modulated Fourier expansion of the numerical solution satisfy*

$$\widehat{\mathcal{I}}_h[\mathbf{z}_h](t) \;=\; \widehat{\mathcal{I}}_h[\mathbf{z}_h](0) \;+\; \mathcal{O}(th^N) \tag{6.25}$$

for $0 \le t \le T$. *Moreover,*

$$\widehat{\mathcal{I}}_h[\mathbf{z}_h](t) \;=\; \sigma(h\omega)\, 2\omega^2\, \|z_{h,1}^1(t)\|^2 + \mathcal{O}(h^2)\,, \tag{6.26}$$

where again $\sigma(h\omega) = \mathrm{sinc}(h\omega)\phi(h\omega)/\psi(h\omega)$. □

Symplectic methods have $\psi(\xi) = \mathrm{sinc}(\xi)\,\phi(\xi)$ and hence $\sigma(h\omega) = 1$. To be able to also treat methods where $\sigma(h\omega)$ can be small, we need to sharpen the estimates of Lemma 6.5. Close scrutiny of the equations (5.20) that determine the coefficient functions of the modulated Fourier expansion, shows that the $\mathcal{O}(h^2)$ term in (6.26) contains a factor $\phi(h\omega)^2$, and that the $\mathcal{O}(th^N)$ term in (6.25) can be put in the form $\mathcal{O}(t\phi(h\omega)^N h^N) + \mathcal{O}(th^{N+m})$ with an arbitrary integer $m \ge 0$; cf. (5.12). Assume now that

$$\phi \text{ is analytic with no real zeros other than integral multiples of } \pi. \tag{6.27}$$

This condition ensures that $|\phi(h\omega)|^2 \ge ch^m$ for some m if $h\omega$ satisfies (5.10). Under the conditions of Theorem 5.2, in particular, (4.1) and (5.10), the improved bounds of the remainder terms yield the following estimates for $\mathcal{I}_h = \widehat{\mathcal{I}}_h/\sigma(h\omega)$:

$$\mathcal{I}_h[\mathbf{z}_h](t) \;=\; \mathcal{I}_h[\mathbf{z}_h](0) \;+\; \mathcal{O}(th^N) \tag{6.28}$$

$$\mathcal{I}_h[\mathbf{z}_h](t) \;=\; 2\omega^2 \,\|z_{h,1}^1(t)\|^2 + \mathcal{O}(h^2)\;. \tag{6.29}$$

Relationship with the Total and the Oscillatory Energy. The almost-invariants

$$\mathcal{I}_h = \frac{1}{\sigma(h\omega)}\,\widehat{\mathcal{I}}_h\;, \quad \mathcal{H}_h = \widehat{\mathcal{H}}_h - \Big(1 - \frac{1}{\sigma(h\omega)}\Big)\widehat{\mathcal{I}}_h \tag{6.30}$$

of the coefficient functions of the modulated Fourier expansion are then close to the total energy H and the oscillatory energy I along the numerical solution $(x_n, \dot{x}_n)$:

Theorem 6.6. *Under the conditions of Theorems 5.2 and condition (6.27),*

$$\begin{aligned}
\mathcal{H}_h[\mathbf{z}_h](t) &= \mathcal{H}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N)\;, \quad & \mathcal{I}_h[\mathbf{z}_h](t) &= \mathcal{I}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N)\\
\mathcal{H}_h[\mathbf{z}_h](t) &= H(x_n,\dot{x}_n) + \mathcal{O}(h)\;, & \mathcal{I}_h[\mathbf{z}_h](t) &= I(x_n,\dot{x}_n) + \mathcal{O}(h)
\end{aligned}$$

holds for $0 \le t = nh \le T$. *The constants symbolized by* $\mathcal{O}$ *depend on* E, N *and* T.

Proof. The upper two relations follow directly from (6.22) and (6.28). Theorems 5.2 and 5.3 show

$$\begin{aligned}
\omega\, x_{n,1} &= \omega\big(e^{i\omega t} z_{h,1}^1(t) + e^{-i\omega t} z_{h,1}^{-1}(t)\big) + \mathcal{O}(h)\\
\dot{x}_{n,1} &= i\omega\big(e^{i\omega t} z_{h,1}^1(t) - e^{-i\omega t} z_{h,1}^{-1}(t)\big) + \mathcal{O}(h)\;.
\end{aligned}$$

With the identity $\|v+\overline{v}\|^2 + \|v-\overline{v}\|^2 = 4\|v\|^2$, this implies

$$I(x_n,\dot{x}_n) \;=\; 2\omega^2\,\|z_{h,1}^1(t)\|^2 + \mathcal{O}(h)\;.$$

A comparison with (6.29) then gives the stated relation between I and $\mathcal{I}_h$. The relation between H and $\mathcal{H}_h$ is proved in the same way, using in addition (6.23). □

XIII.7 Long-Time Near-Conservation of Total and Oscillatory Energy

With the results of the previous section, we can now show that the numerical method nearly preserves the total energy H and the oscillatory energy I over time intervals of length $C_N h^{-N+1}$, for any N for which the non-resonance condition (5.10) is satisfied. Such a result is due to Hairer & Lubich (2000a).
For convenience we restate the assumptions:

- the energy bound (2.3): $\frac{1}{2}\|\dot{x}(0)\|^2 + \frac{1}{2}\|\Omega x(0)\|^2 \le E$;
- the condition on the numerical solution: the values Φx_n stay in a compact subset of a domain on which the potential U is smooth;

- the conditions on the filter functions: ψ and ϕ are even, real-analytic, and have no real zeros other than integral multiples of π; they satisfy $\psi(0) = \phi(0) = 1$ and (4.1):

$$\begin{aligned} &|\psi(h\omega)| \,\leq\, C_1 \,\mathrm{sinc}^2(\tfrac{1}{2}h\omega)\,, \qquad |\phi(h\omega)| \,\leq\, C_2\, |\mathrm{sinc}(\tfrac{1}{2}h\omega)|\,, \\ &|\psi(h\omega)\phi(h\omega)| \,\leq\, C_3\, |\mathrm{sinc}(h\omega)|\,; \end{aligned} \tag{7.1}$$

- the condition $h\omega \geq c_0 > 0$;
- the non-resonance condition (5.10): for some $N \geq 2$,

$$|\sin(\tfrac{1}{2}kh\omega)| \,\geq\, c\,\sqrt{h} \qquad \text{for} \quad k = 1, \ldots, N.$$

Theorem 7.1. *Under the above conditions, the numerical solution of (2.1) obtained by the method (2.7)–(2.8) with (2.9) satisfies*

$$\begin{aligned} H(x_n, \dot{x}_n) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I(x_n, \dot{x}_n) &= I(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \qquad \textit{for} \quad 0 \leq nh \leq h^{-N+1}.$$

The constants symbolized by $\mathcal{O}$ are independent of n, h, ω satisfying the above conditions, but depend on N and the constants in the conditions.

Proof. The estimates of Theorem 6.6 hold uniformly over bounded intervals. We now apply those estimates repeatedly on intervals of length h, for modulated Fourier expansions corresponding to different starting values. As long as $(x_n, \dot{x}_n)$ satisfies the bounded-energy condition (2.3) (possibly with a larger constant E), Theorem 5.2 gives a modulated Fourier expansion that corresponds to starting values $(x_n, \dot{x}_n)$. We denote the vector of coefficient functions of this expansion by $\mathbf{z}_n(t)$:

$$\mathbf{z}_n = (z_n^{-N+1}, \ldots, z_n^{-1}, y_n, z_n^1, \ldots, z_n^{N-1})$$

(omitting the notational dependence on h for simplicity). Because of the uniqueness, up to $\mathcal{O}(h^{N+1})$, of the coefficient functions of the modulated Fourier expansion constructed by (5.20), the following diagram commutes up to terms of size $\mathcal{O}(h^{N+1})$:

$$\begin{array}{ccc} (x_n, \dot{x}_n) & \longleftrightarrow & (\mathbf{z}_n(0), \dot{\mathbf{z}}_n(0)) \\ & & \Big\downarrow \text{flow} \\ \Big\downarrow \begin{array}{l}\text{numerical}\\ \text{method}\end{array} & & (\mathbf{z}_n(h), \dot{\mathbf{z}}_n(h)) \\ & & = \ (\text{up to } \mathcal{O}(h^{N+1})) \\ (x_{n+1}, \dot{x}_{n+1}) & \longleftrightarrow & (\mathbf{z}_{n+1}(0), \dot{\mathbf{z}}_{n+1}(0)) \end{array}$$

The construction of the coefficient functions via (5.20) shows that also higher derivatives of $\mathbf{z}_n$ at h and $\mathbf{z}_{n+1}$ at 0 differ by only $\mathcal{O}(h^{N+1})$. From the above diagram and Theorem 6.6 we thus obtain

$$\begin{aligned}\mathcal{H}_h[\mathbf{z}_{n+1}](0) &= \mathcal{H}_h[\mathbf{z}_n](h) + \mathcal{O}(h^{N+1}) \\ &= \mathcal{H}_h[\mathbf{z}_n](0) + \mathcal{O}(h^{N+1}) \,.\end{aligned}$$

Repeated use of this relation gives

$$\mathcal{H}_h[\mathbf{z}_n](0) = \mathcal{H}_h[\mathbf{z}_0](0) + \mathcal{O}(nh^{N+1}) \,.$$

Moreover, by Theorem 6.6 the coefficient functions corresponding to the starting values $(x_n, \dot{x}_n)$ and $(x_0, \dot{x}_0)$ satisfy

$$\begin{aligned}\mathcal{H}_h[\mathbf{z}_n](0) &= H(x_n, \dot{x}_n) + \mathcal{O}(h) \,, \\ \mathcal{H}_h[\mathbf{z}_0](0) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \,.\end{aligned}$$

So we obtain

$$\begin{aligned}H(x_n, \dot{x}_n) - H(x_0, \dot{x}_0) &= \mathcal{H}_h[\mathbf{z}_n](0) - \mathcal{H}_h[\mathbf{z}_0](0) + \mathcal{O}(h) \\ &= \mathcal{O}(nh^{N+1}) + \mathcal{O}(h) \,,\end{aligned}$$

which gives the desired bound for the deviation of the total energy along the numerical solution. The same argument applies to $I(x_n, \dot{x}_n)$. $\square$

The imposed bounds of ψ and ϕ become important when $h\omega$ is close to an integral multiple of π. Are these conditions also sufficient to guarantee favourable energy behaviour uniformly in $h\omega$, arbitrarily close to multiples of π? Unfortunately the answer is negative (see Fig. 2.5 to Fig. 2.7). The analysis of method (2.7)–(2.9) for exact resonances $h\omega = m\pi$ with integer m shows that stronger conditions

$$|\psi(h\omega)| \le C\,|\mathrm{sinc}(h\omega)| \,, \qquad |\psi(h\omega)\phi(h\omega)| \le C\,\mathrm{sinc}^2(h\omega) \tag{7.2}$$

are required. Even this is not sufficient for near-conservation of the total and the oscillatory energy for $h\omega$ *near* a multiple of π. For linear problems

$$\ddot{x} + \begin{pmatrix} 0 & 0 \\ 0 & \omega^2 \end{pmatrix} x = -Ax$$

with a two-dimensional symmetric matrix A with $a_{00} > 0$, and with initial values satisfying the bounded-energy condition (2.3), Hairer & Lubich (2000a) show that the numerical method conserves the total energy up to $\mathcal{O}(h)$ uniformly for all times and for all values of $h\omega$, if and only if

$$\psi(\xi) = \mathrm{sinc}^2(\xi)\,\phi(\xi) \,. \tag{7.3}$$

There is *no* method (2.7)-(2.8) which approximately preserves the oscillatory energy I uniformly for all $h\omega$ in a fixed open interval that contains a multiple of 2π.

In summary, the bad effect of step-size resonances on the energy behaviour of the method cannot be eliminated, but it can be considerably mitigated by an appropriate choice of the filter functions ψ and ϕ.

XIII.8 Energy Behaviour of the Störmer–Verlet Method

The results of Sections XIII.5–XIII.7 provide new insight into the energy behaviour of the classical Störmer–Verlet method. We present in this section weakened versions of results of Hairer & Lubich (2000b).

In applications, the Störmer–Verlet method is typically used with step sizes h for which the product with the highest frequency ω is in the range of linear stability, but is bounded away from 0. For example, in spatially discretized wave equations, $h\omega$ is known as the CFL number, which is typically kept near 1. Values of $h\omega$ around $\frac{1}{2}$ are often used in molecular dynamics. In contrast, the backward error analysis of Chap. IX explains the long-time energy behaviour only for $h\omega \to 0$.

Consider now applying the Störmer–Verlet method to the nonlinear model problem (2.1)-(2.3),

$$x_{n+1} - 2x_n + x_{n-1} \;=\; -\,h^2\Omega^2 x_n - h^2\nabla U(x_n)\,, \tag{8.1}$$

with $h\omega < 2$ for linear stability. The method is made accessible to the analysis of Sections XIII.3–XIII.7 by rewriting it as a trigonometric method (2.6) with a *modified frequency*:

$$x_{n+1} - 2\cos(h\widetilde{\Omega})\,x_n + x_{n-1} \;=\; -\,h^2\nabla U(x_n)\,, \tag{8.2}$$

where

$$\widetilde{\Omega} = \begin{pmatrix} 0 & 0 \\ 0 & \widetilde{\omega} I \end{pmatrix} \qquad \text{with} \quad \sin(\tfrac{1}{2}h\widetilde{\omega}) = \tfrac{1}{2}h\omega\,. \tag{8.3}$$

The velocity approximation

$$\dot{x}_n \;=\; \frac{x_{n+1} - x_{n-1}}{2h}$$

does not correspond to the velocity approximation (2.11) of the trigonometric method, but this presents only a minor technical difficulty. We show that the following *modified energies* are well conserved by the Störmer–Verlet method:

$$\begin{aligned} H^*(x,\dot{x}) &\;=\; H(x,\dot{x}) + \tfrac{1}{2}\gamma\,\|\dot{x}_1\|^2 \\ I^*(x,\dot{x}) &\;=\; I(x,\dot{x}) \;+ \tfrac{1}{2}\gamma\,\|\dot{x}_1\|^2 \end{aligned} \qquad \text{with} \quad \gamma = \frac{1}{1-\frac{1}{4}(h\omega)^2} - 1\,. \tag{8.4}$$

Here H and I are again the total and the oscillatory energy of the system (2.1) (defined with the original ω, not with $\widetilde{\omega}$).

Theorem 8.1. *Let the Störmer–Verlet method be applied to the problem (2.1)-(2.3) with a step size h for which $0 < c_0 \le h\omega \le c_1 < 2$ and $|\sin(\frac{1}{2}kh\widetilde{\omega})| \ge c\sqrt{h}$ for $k = 1,\ldots,N$ for some $N \ge 2$ and $c > 0$. Suppose further that the numerical solution values x_n stay in a region on which all derivatives of U are bounded. Then, the modified energies along the numerical solution satisfy*

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= H^*(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I^*(x_n, \dot{x}_n) &= I^*(x_0, \dot{x}_0) \ + \mathcal{O}(h) \end{aligned} \qquad \textit{for} \quad 0 \le nh \le h^{-N+1} \, . \tag{8.5}$$

The constants symbolized by $\mathcal{O}$ are independent of n, h, ω with the above conditions.

Proof. With the modified velocities x_n' defined by

$$2h \operatorname{sinc}(h\widetilde{\Omega})\, x_n' \;=\; x_{n+1} - x_{n-1}$$

method (8.2) becomes a method (2.6) with (2.11), or equivalently (2.7)-(2.8), with $\widetilde{\omega}$ instead of ω and with $\psi(\xi) = \phi(\xi) = 1$.

The condition $0 < c_0 \le h\omega \le c_1 < 2$ implies $|\sin(\frac{1}{2}kh\widetilde{\omega})| \ge c_2 > 0$ for $k = 1, 2$, and hence conditions (7.1) are trivially satisfied with $h\widetilde{\omega}$ instead of $h\omega$. We are thus in the position to apply Theorem 7.1, which yields

$$\begin{aligned} \widetilde{H}(x_n, x_n') &= \widetilde{H}(x_0, x_0') + \mathcal{O}(h) \\ \widetilde{I}(x_n, x_n') &= \widetilde{I}(x_0, x_0') + \mathcal{O}(h) \end{aligned} \qquad \text{for } \; 0 \le nh \le h^{-N+1} \, , \tag{8.6}$$

where $\widetilde{H}$ and $\widetilde{I}$ are defined in the same way as H and I, but with $\widetilde{\omega}$ in place of ω. The components of the Störmer–Verlet velocities $\dot{x}_n$ and the modified velocities x_n' are related by

$$\dot{x}_{n,0} = x_{n,0}' \, , \qquad \dot{x}_{n,1} = \operatorname{sinc}(h\widetilde{\omega})\, x_{n,1}' = \frac{\omega}{\widetilde{\omega}} \sqrt{1 - \tfrac{1}{4}h^2\omega^2}\; x_{n,1}' \, , \tag{8.7}$$

so that

$$\begin{aligned} \widetilde{I}(x_n, x_n') &= \frac{1}{2}\|x_{n,1}'\|^2 + \frac{1}{2}\widetilde{\omega}^2 \|x_{n,1}\|^2 \\ &= \frac{1}{2}\,\frac{\widetilde{\omega}^2}{\omega^2}\,\frac{1}{1 - \frac{1}{4}h^2\omega^2}\;\|\dot{x}_{n,1}\|^2 + \frac{1}{2}\,\frac{\widetilde{\omega}^2}{\omega^2}\,\omega^2 \|x_{n,1}\|^2 \\ &= \frac{\widetilde{\omega}^2}{\omega^2}\, I^*(x_n, \dot{x}_n) \, . \end{aligned} \tag{8.8}$$

Similarly,

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= \frac{1}{2}\|\dot{x}_{n,0}\|^2 + U(x_n) + I^*(x_n, \dot{x}_n) \\ &= \widetilde{H}(x_n, x_n') + \left(\frac{\omega^2}{\widetilde{\omega}^2} - 1\right) \widetilde{I}(x_n, x_n') \, , \end{aligned} \tag{8.9}$$

and hence (8.6) yields the result. □

For fixed $h\omega \ge c_0 > 0$ and $h \to 0$, the maximum deviation in the energy does not tend to 0, due to the highly oscillatory term $\frac{1}{2}\gamma\|\dot{x}_1\|^2$ in $H^*(x, \dot{x})$ and $I^*(x, \dot{x})$. We show, however, that *time averages* of H and I are nearly preserved over long time. For an arbitrary fixed $T > 0$, consider the averages over intervals of length T,

$$\begin{aligned} \overline{H}_n &= \frac{1}{T}\, h \sum_{|jh|\le T/2} H(x_{n+j}, \dot{x}_{n+j}) \\ \overline{I}_n &= \frac{1}{T}\, h \sum_{|jh|\le T/2} I(x_{n+j}, \dot{x}_{n+j}) \,. \end{aligned} \tag{8.10}$$

Theorem 8.2. *Under the conditions of Theorem 8.1, the time averages of the total and the oscillatory energy along the numerical solution satisfy*

$$\begin{aligned} \overline{H}_n &= \overline{H}_0 + \mathcal{O}(h) \\ \overline{I}_n &= \overline{I}_0 \ + \mathcal{O}(h) \end{aligned} \qquad \textit{for} \quad 0 \le nh \le h^{-N+1} \,. \tag{8.11}$$

The constants symbolized by $\mathcal{O}$ are independent of n, h, ω with the above conditions.

Proof. We show

$$\begin{aligned} \overline{H}_n &= H^*(x_n, \dot{x}_n) - \frac{1}{2}\frac{\gamma}{1+\gamma}\, I^*(x_n, \dot{x}_n) + \mathcal{O}(h) \\ \overline{I}_n &= I^*(x_n, \dot{x}_n) - \frac{1}{2}\frac{\gamma}{1+\gamma}\, I^*(x_n, \dot{x}_n) + \mathcal{O}(h) \,, \end{aligned} \tag{8.12}$$

which implies the result by Theorem 8.1. Consider the modulated Fourier expansions of x_n and x_n' for $t = nh$ in a bounded interval. Theorem 5.3 shows that

$$x_{n,1}' = i\widetilde{\omega}\big(e^{i\widetilde{\omega}t} z_{h,1}^1(t) - e^{-i\widetilde{\omega}t}\overline{z_{h,1}^1}(t)\big) + \mathcal{O}(h) \,, \qquad t = nh \,,$$

with $z_{h,1}^1(t)$ from the modulated Fourier expansion of Theorem 5.2 (with $\widetilde{\omega}$ instead of ω). With (8.7) it follows that

$$\dot{x}_{n,1} = i\omega\sqrt{1 - \tfrac{1}{4}h^2\omega^2}\, \big(e^{i\widetilde{\omega}t} z_{h,1}^1(t) - e^{-i\widetilde{\omega}t}\overline{z_{h,1}^1}(t)\big) + \mathcal{O}(h) \,,$$

and therefore, recalling the definition of γ,

$$\|\dot{x}_{n,1}\|^2 = \omega^2 \frac{1}{1+\gamma}\, \Big(2\,\|z_{h,1}^1(t)\|^2 - 2\,\mathrm{Re}\, e^{2i\widetilde{\omega}t} z_{h,1}^1(t)^2\Big) + \mathcal{O}(h) \,.$$

Theorems 5.2 and 5.3 yield

$$2\widetilde{\omega}^2 \|z_{h,1}^1(t)\|^2 = \widetilde{I}(x_n, x_n') + \mathcal{O}(h)$$

and hence, by (8.8),

$$2\omega^2 \|z_{h,1}^1(t)\|^2 = I^*(x_n, \dot{x}_n) + \mathcal{O}(h) \,.$$

A partial summation shows that the time average over the highly oscillatory terms $e^{2i\widetilde{\omega}t}\omega^2 z_{h,1}^1(t)^2$ is $\mathcal{O}(h)$. This finally gives

$$\frac{1}{T}\, h \sum_{|j|\le T/2} \|\dot{x}_{j,1}\|^2 = \frac{1}{1+\gamma}\, I^*(x_n, \dot{x}_n) + \mathcal{O}(h) \,.$$

Taking the time averages in the expressions of the definition (8.4) of H^* and I^* then yields (8.12). □

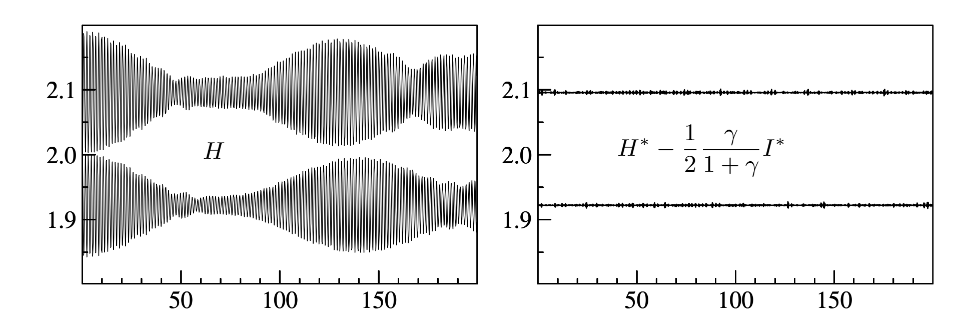


Fig. 8.1. Total energies (left) and their predicted averages (right) for the Störmer–Verlet method and for two different initial values, with $\omega = 50$ and h such that $h\omega = 0.8$

Figure 8.1 illustrates the above result. It shows the total energy H for two different initial values on the left, and the averages as predicted by the expression on the right-hand side of (8.12) on the right picture. The initial values are as in Chap. I with the exception of $x_{1,1}(0)$ and $\dot{x}_{1,1}(0)$. We take $x_{1,1}(0) = \sqrt{2}/\omega$, $\dot{x}_{1,1}(0) = 0$ for one set of initial values and $x_{1,1}(0) = 0$, $\dot{x}_{1,1}(0) = \sqrt{2}$ for the other. The total energies at the initial values are 2.00240032 and 2, respectively.

XIII.9 Systems with Several Constant Frequencies

This section studies the conservation of invariants and almost-invariants along numerical approximations of an extension of (2.1) to systems with the Hamiltonian function

$$H(p,q) = \frac{1}{2}\,p^T M^{-1} p + \frac{1}{2\varepsilon^2}\,q^T A q + U(q) \tag{9.1}$$

with a positive definite constant matrix M and a positive semi-definite constant matrix A. With the Cholesky decomposition $M = LL^T$ and the canonical transformation $\widetilde{p} = L^{-1}p$, $\widetilde{q} = L^T q$ we obtain a Hamiltonian where the mass matrix is the identity matrix and A is transformed to $\widetilde{A} = L^{-1}AL^{-T}$. Diagonalizing $\widetilde{A} = Q\Lambda Q^T$ and transforming to $x = Q^T\widetilde{q}$ then yields a Hamiltonian of the form (we omit the tilde on $\widetilde{U}(x) = U(q)$ and $\widetilde{H}(x,\dot{x}) = H(p,q)$)

$$H(x,\dot{x}) = \frac{1}{2}\sum_{j=0}^{\ell}\Big(\|\dot{x}_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2}\,\|x_j\|^2\Big) + U(x), \tag{9.2}$$

where $x = (x_0, x_1, \ldots, x_\ell)$ with $x_j \in \mathbb{R}^{d_j}$, $\lambda_0 = 0$, and $\lambda_j > 0$ for $j \geq 1$ are all distinct. After rescaling ε we may assume $\lambda_j \geq 1$ for $j = 1, \ldots, \ell$.

Following Cohen, Hairer & Lubich (2004) we extend the results of the previous sections to the multi-frequency case $\ell > 1$. Modulated Fourier expansions are again the basic analytical tool. A new aspect is possible resonance among the λ_j.

XIII.9.1 Oscillatory Energies and Resonances

The equations of motion for the Hamiltonian system (9.2) can be written as the system of second-order differential equations

$$\ddot{x} = -\Omega^2 x + g(x), \tag{9.3}$$

where $\Omega = \operatorname{diag}(\omega_j I)$ with the frequencies $\omega_j = \lambda_j/\varepsilon$ and $g(x) = -\nabla U(x)$. As suitable numerical methods we consider again the class of trigonometric integrators studied in Sect. XIII.2, (2.6) with (2.11), with filter functions ψ and ϕ.

We are interested in the long-time near-conservation of the total energy $H(x, \dot{x})$ and the oscillatory energies

$$I_j(x, \dot{x}) = \frac{1}{2}\Big(\|\dot{x}_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2}\,\|x_j\|^2\Big) \qquad \text{for} \quad j \geq 1 \tag{9.4}$$

or suitable linear combinations thereof. Benettin, Galgani & Giorgilli (1989) have shown that the quantities

$$I_\mu(x, \dot{x}) = \sum_{j=1}^{\ell} \frac{\mu_j}{\lambda_j}\, I_j(x, \dot{x}) \tag{9.5}$$

are approximately preserved along every bounded solution of the Hamiltonian system that has a total energy bounded independently of ε, on exponentially long time intervals of size $\mathcal{O}(e^{c/\varepsilon})$ if the potential $U(x)$ is analytic and $\mu = (\mu_1, \ldots, \mu_\ell)$ is orthogonal to the *resonance module*

$$\mathcal{M} = \{k \in \mathbb{Z}^\ell : k_1\lambda_1 + \ldots + k_\ell\lambda_\ell = 0\}, \tag{9.6}$$

if a diophantine non-resonance condition holds outside $\mathcal{M}$. (Cf. also Sect. XIII.9.4 below.)

Since $\mu = \lambda$ is orthogonal to $\mathcal{M}$, the total oscillatory energy $\sum_{j=1}^{\ell} I_j(x, \dot{x})$ of the system is approximately preserved independently of the resonance module $\mathcal{M}$. Subtracting this expression from the total energy (1.7), we see that also the *smooth energy*

$$K(x, \dot{x}) = \frac{1}{2}\|\dot{x}_0\|^2 + U(x) \tag{9.7}$$

is approximately preserved. With an ε-independent bound of the total energy $H(x, \dot{x})$ we have $x_j = \mathcal{O}(\varepsilon)$ for $j = 1, \ldots, \ell$, so that $K(x, \dot{x})$ is close to the Hamiltonian of the reduced system in which all oscillatory degrees of freedom are taken out, $H_0(x_0, \dot{x}_0) = \frac{1}{2}\|\dot{x}_0\|^2 + U(x_0, 0, \ldots, 0)$.

Example 9.1. To illustrate the conservation of the various energies, we consider a Hamiltonian (1.7) with $\ell = 3$, $\lambda = (1, \sqrt{2}, 2)$ and we assume that the dimensions of x_j are all 1 with the exception of that of $x_1 = (x_{1,1}, x_{1,2})$ which is 2. The resonance module is then $\mathcal{M} = \{(k_1, 0, k_3)\,;\, k_1 + 2k_3 = 0\}$. We take $\varepsilon^{-1} = \omega = 70$, the potential

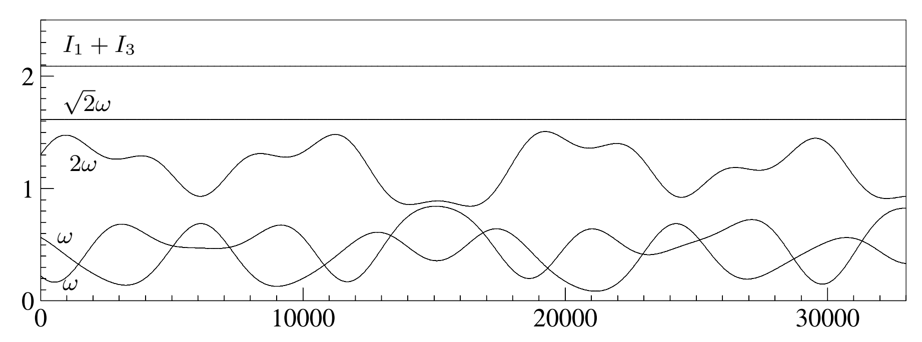


Fig. 9.1. Oscillatory energies of the individual components (the frequencies $\lambda_j \omega = \lambda_j/\varepsilon$ are indicated) and the sum $I_1 + I_3$ of the oscillatory energies corresponding to the resonant frequencies ω and 2ω

$$U(x) = (0.05 + x_{1,1} + x_{1,2} + x_2 + 2.5\, x_3)^4 + \frac{1}{8} x_0^2\, x_{1,1}^2 + \frac{1}{2} x_0^2, \tag{9.8}$$

and $x(0) = (1, 0.3\varepsilon, 0.8\varepsilon, -1.1\varepsilon, 0.7\varepsilon)$, $\dot{x}(0) = (-0.2, 0.6, 0.7, -0.9, 0.8)$ as initial values. We consider I_μ for $\mu = (1, 0, 2)$ and $\mu = (0, \sqrt{2}, 0)$, which are both orthogonal to $\mathcal{M}$. In Fig. 9.1 we plot the oscillatory energies for the individual components of the system. The corresponding frequencies are attached to the curves. We also plot the sum $I_1 + I_3$ of the three oscillatory energies corresponding to the resonant frequencies $1/\varepsilon$ and $2/\varepsilon$. We see that $I_1 + I_3$ as well as I_2 (which are I_μ for the above two vectors $\mu \perp \mathcal{M}$) are well conserved over long times up to small oscillations of size $\mathcal{O}(\varepsilon)$. There is an energy exchange between the two components corresponding to the same frequency $1/\varepsilon$, and on a larger scale an energy exchange between I_1 and I_3.

Numerical Experiment. As a first method we take (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \operatorname{sinc}(\xi)$, and we apply it with large step sizes so that $h\omega = h/\varepsilon$ takes the values 1, 2, 4, and 8. Figure 9.2 shows the various oscillatory energies which can be compared to the exact values in Fig. 9.1. For all step sizes, the oscillatory energy corresponding to the frequency $\sqrt{2}\omega$ and the sum $I_1 + I_3$ are well conserved on long time intervals. Oscillations in these expressions increase with h. The energy exchange between resonant frequencies is close to that of the exact solution. We have not plotted the total energy $H(x_n, \dot{x}_n)$ nor the smooth energy $K(x_n, \dot{x}_n)$ of (9.7). Both are well conserved over long times.

We repeat this experiment with the method where $\phi(\xi) = 1$ and $\psi(\xi) = \operatorname{sinc}^2(\xi/2)$ (Fig. 9.3). Only the oscillatory energy corresponding to $\sqrt{2}\omega$ is approximately conserved over long times. Neither the expression $I_1 + I_3$ nor the total energy (not shown) are conserved. The smooth energy $K(x_n, \dot{x}_n)$ is, however, well conserved.

Figure 9.4 shows the corresponding result for the method with $\phi(\xi) = \operatorname{sinc}(\xi)$ and $\psi(\xi) = \operatorname{sinc}(\xi)\phi(\xi)$. The oscillatory energy for $\sqrt{2}\omega$ and also $I_1 + I_3$ are well conserved. However, the energy exchange between the resonant frequencies is not correctly reproduced.

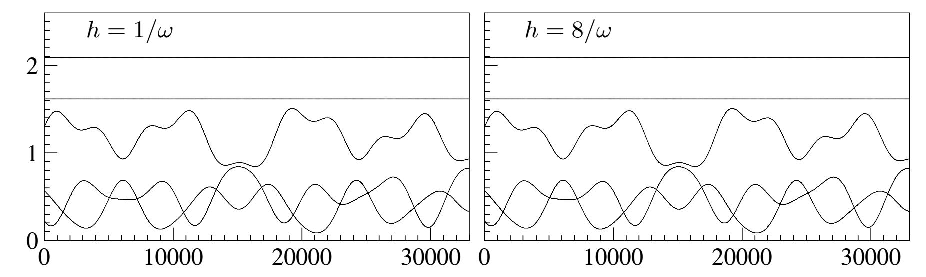


Fig. 9.2. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \mathrm{sinc}(\xi)$

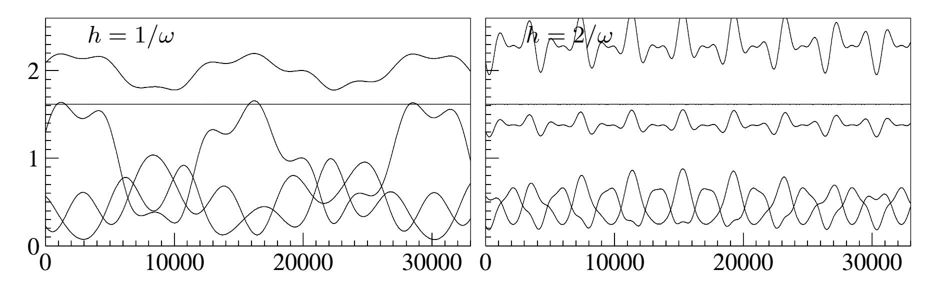


Fig. 9.3. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \mathrm{sinc}^2(\xi/2)$

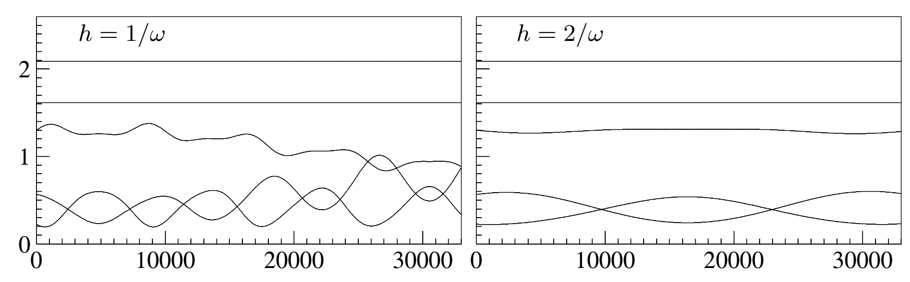


Fig. 9.4. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = \mathrm{sinc}(\xi)$ and $\psi(\xi) = \mathrm{sinc}(\xi)\phi(\xi)$

XIII.9.2 Multi-Frequency Modulated Fourier Expansions

The above numerical phenomena can be understood with a multi-frequency version of the modulated Fourier expansions studied in the previous chapter. We only outline the derivation and properties, since they are in large parts similar to the single-frequency case. More details can be found in Cohen, Hairer & Lubich (2004). We assume conditions that extend those of the previous sections:

- The energy of the initial values is bounded independently of ε,

$$\frac{1}{2}\|\dot{x}(0)\|^2 + \frac{1}{2}\|\Omega x(0)\|^2 \leq E \,. \tag{9.9}$$

- The numerical solution values Φx_n stay in a compact subset of a domain on which the potential U is smooth.
- We impose a lower bound on the step size: $h/\varepsilon \geq c_0 > 0$.
- We assume the numerical non-resonance condition

$$\Big|\sin\Big(\frac{h}{2\varepsilon}\,k\cdot\lambda\Big)\Big| \;\geq\; c\sqrt{h} \quad \text{for all } k\in\mathbb{Z}^\ell\setminus\mathcal{M} \text{ with } |k|\leq N, \tag{9.10}$$

  for some $N\geq 2$ and $c>0$.
- For the filter functions we assume that for $\xi_j = h\lambda_j/\varepsilon$ $(j=1,\ldots,\ell)$,

$$\begin{aligned} |\psi(\xi_j)| &\leq C_1\,\mathrm{sinc}^2(\tfrac12\xi_j)\,,\\ |\phi(\xi_j)| &\leq C_2\,|\mathrm{sinc}(\tfrac12\xi_j)|\,,\\ |\psi(\xi_j)| &\leq C_3\,|\mathrm{sinc}(\xi_j)\,\phi(\xi_j)|\,. \end{aligned} \tag{9.11}$$

The conditions on the filter functions are somewhat stronger than necessary, but they facilitate the presentation in the following.

For a given vector $\lambda=(\lambda_1,\ldots,\lambda_\ell)$ and for the resonance module $\mathcal{M}$ defined by (9.6), we let $\mathcal{K}$ be a set of representatives of the equivalence classes in $\mathbb{Z}^\ell/\mathcal{M}$ which are chosen such that for each $k\in\mathcal{K}$ the sum $|k|=|k_1|+\ldots+|k_\ell|$ is minimal in the equivalence class $[k]=k+\mathcal{M}$, and with $k\in\mathcal{K}$, also $-k\in\mathcal{K}$. We denote, for N of (6.3),

$$\mathcal{N}=\{k\in\mathcal{K}:|k|<N\},\qquad \mathcal{N}^*=\mathcal{N}\setminus\{(0,\ldots,0)\}. \tag{9.12}$$

The following multi-frequency version of Theorem XIII.5.2 establishes a modulated Fourier expansion for the numerical solution.

Theorem 9.2. *Consider the numerical solution of the system (9.3) by the method (2.6) with step size h. Under conditions (9.9)–(9.11), the numerical solution admits an expansion*

$$x_n \;=\; y(t)+\sum_{k\in\mathcal{N}^*} e^{ik\cdot\omega t}z^k(t)+\Psi\cdot\mathcal{O}(t^2h^N) \tag{9.13}$$

with $\omega=\lambda/\varepsilon$*, uniformly for* $0\leq t=nh\leq T$ *and* ε *and* h *satisfying* $h/\varepsilon\geq c_0>0$*. The modulation functions together with all their derivatives (up to some arbitrarily fixed order) are bounded by*

$$\begin{aligned} y_0 &= \mathcal{O}(1), & y_j &= \mathcal{O}(\varepsilon^2)\\ z_j^{\pm\langle j\rangle} &= \mathcal{O}(\varepsilon), & \dot z_j^{\pm\langle j\rangle} &= \mathcal{O}(\varepsilon^2)\\ z_j^k &= \mathcal{O}(h\varepsilon^{|k|}) & &\text{for } k\neq\pm\langle j\rangle \end{aligned} \tag{9.14}$$

for $j=1,\ldots,\ell$*. Here,* $\langle j\rangle=(0,\ldots,1,\ldots,0)$ *is the* j*th unit vector. The last estimate holds also for* z_0^k *for all* $k\in\mathcal{N}^*$*. Moreover, the function* y *is real-valued and* $z^{-k}=\overline{z^k}$ *for all* $k\in\mathcal{N}^*$*. The constants symbolized by the* $\mathcal{O}$*-notation are independent of* h*,* ε *and* λ_j *with (9.10), but they depend on* E*,* N*,* c*, and* T*.*

The proof extends that of Theorem XIII.5.2. In terms of the difference operator of the method, $L(hD) = e^{hD} - 2\cos h\Omega + e^{-hD}$, the functions $y(t)$ and $z^k(t)$ are constructed such that, up to terms of size $\Psi \cdot \mathcal{O}(h^{N+2})$,

$$L(hD)y = h^2\Psi\Big(g(\Phi y) + \sum_{s(\alpha)\sim 0} \frac{1}{m!}\, g^{(m)}(\Phi y)(\Phi z)^\alpha\Big)$$
$$L(hD + ihk\cdot\omega)z^k = h^2\Psi \sum_{s(\alpha)\sim k} \frac{1}{m!}\, g^{(m)}(\Phi y)(\Phi z)^\alpha.$$

Here, the sums on the right-hand side are over all $m \geq 1$ and over multi-indices $\alpha = (\alpha_1, \ldots, \alpha_m)$ with $\alpha_j \in \mathcal{N}^*$, for which the sum $s(\alpha) = \sum_{j=1}^m \alpha_j$ satisfies the relation $s(\alpha) \sim k$, that is, $s(\alpha) - k \in \mathcal{M}$. The notation $(\Phi z)^\alpha$ is short for the m-tuple $(\Phi z^{\alpha_1}, \ldots, \Phi z^{\alpha_m})$.

A similar expansion to that for x_n exists also for the velocity approximation $\dot{x}_n$, like in Theorem XIII.5.3. As a consequence, the oscillatory energy (9.4) along the numerical solution takes the form, at $t = nh \leq T$,

$$I_j(x_n, \dot{x}_n) = 2\omega_j^2\, \|z_j^{\langle j\rangle}(t)\|^2 + \mathcal{O}(\varepsilon). \tag{9.15}$$

With the first terms of the modulated Fourier expansion one proves, as in Theorems XIII.4.1 and XIII.4.2, error bounds over bounded time intervals which are of second order in the positions and of first order in the velocities:

$$\|x_n - x(t_n)\| \leq C\,h^2\,, \qquad \|\dot{x}_n - \dot{x}(t_n)\| \leq C\,h\,, \tag{9.16}$$

where C is independent of ε, h and n with $nh \leq T$ and of bounds of solution derivatives.

XIII.9.3 Almost-Invariants of the Modulation System

With $y^0(t) = z^0(t) = y(t)$ and $y^k(t) = e^{ik\cdot\omega t} z^k(t)$ for $k \in \mathcal{N}$, where y and z^k are the modulation functions of Theorem 9.2, we denote

$$\mathbf{y} = (y^k)_{k\in\mathcal{N}}, \qquad \mathbf{z} = (z^k)_{k\in\mathcal{N}}.$$

We introduce the extended potential

$$\mathcal{U}(\mathbf{y}) \;=\; U(\Phi y^0) + \sum_{s(\alpha)\sim 0} \frac{1}{m!}\, U^{(m)}(\Phi y^0)(\Phi\mathbf{y})^\alpha\,, \tag{9.17}$$

where the sum is again taken over all $m \geq 1$ and all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_m)$ with $\alpha_j \in \mathcal{N}^*$ for which $s(\alpha) = \sum_j \alpha_j \in \mathcal{M}$. The functions $y^k(t)$ then satisfy

$$\Psi^{-1}\Phi\, h^{-2} L(hD)\, y^k \;=\; -\nabla_{-k}\,\mathcal{U}(\mathbf{y}) + \Phi\cdot\mathcal{O}(h^N), \tag{9.18}$$

where ∇_{-k} denotes the gradient with respect to the variable y^{-k}. This system has almost-invariants that are related to the Hamiltonian H and the oscillatory energies I_μ with $\mu \perp \mathcal{M}$.

The Energy-Type Almost-Invariant of the Modulation System. We multiply (9.18) by $(\dot{y}^{-k})^T$ and sum over $k \in \mathcal{N}$ to obtain

$$\sum_{k\in\mathcal{N}} (\dot{y}^{-k})^T \Psi^{-1}\Phi\, h^{-2} L(hD) y^k + \frac{d}{dt}\mathcal{U}(\mathbf{y}) \;=\; \mathcal{O}(h^N).$$

Since we know bounds of the modulation functions z^k and of their derivatives from Theorem 9.2, we rewrite this relation in terms of the quantities z^k:

$$\sum_{k\in\mathcal{N}} (\dot{z}^{-k} - ik\cdot\omega z^{-k})^T \Psi^{-1}\Phi h^{-2} L(hD + ihk\cdot\omega) z^k + \frac{d}{dt}\mathcal{U}(\mathbf{z}) = \mathcal{O}(h^N). \tag{9.19}$$

As in (6.21) we obtain that the left-hand side of (9.19) can be written as the time derivative of a function $\mathcal{H}^*[\mathbf{z}](t)$ which depends on the values at t of the modulation-function vector $\mathbf{z}$ and its first N time derivatives. The relation (9.19) thus becomes

$$\frac{d}{dt}\mathcal{H}^*[\mathbf{z}](t) \;=\; \mathcal{O}(h^N).$$

Together with the estimates of Theorem 9.2 this construction of $\mathcal{H}^*$ yields the following multi-frequency extension of Lemma XIII.6.4.

Lemma 9.3. *Under the assumptions of Theorem 9.2, the modulation functions* $\mathbf{z} = (z^k)_{k\in\mathcal{N}}$ *of the numerical solution satisfy*

$$\mathcal{H}^*[\mathbf{z}](t) \;=\; \mathcal{H}^*[\mathbf{z}](0) \;+\; \mathcal{O}(th^N) \tag{9.20}$$

for $0 \le t \le T$*. Moreover, at* $t = nh$ *we have*

$$\mathcal{H}^*[\mathbf{z}](t) \;=\; H^*(x_n, \dot{x}_n) + \mathcal{O}(h), \tag{9.21}$$

where, with $\sigma(\xi) = \mathrm{sinc}(\xi)\phi(\xi)/\psi(\xi)$ *and* $\xi_j = h\lambda_j/\varepsilon$*,*

$$H^*(x,\dot{x}) = H(x,\dot{x}) + \sum_{j=1}^{\ell} \bigl(\sigma(\xi_j) - 1\bigr)\, I_j(x,\dot{x}). \tag{9.22}$$

The Momentum-Type Almost-Invariants of the Modulation System. The equations (9.18) have further almost-invariants that result from invariance properties of the extended potential $\mathcal{U}$, similarly as the conservation of angular momentum results from an invariance of the potential U in a mechanical system by Noether's theorem. For $\mu \in \mathbb{R}^\ell$ and $\mathbf{y} = (y^k)_{k\in\mathcal{N}}$ we set

$$S_\mu(\tau)\mathbf{y} = (e^{ik\cdot\mu\tau} y^k)_{k\in\mathcal{N}}, \qquad \tau \in \mathbb{R}$$

so that, by the multi-linearity of the derivative, the definition (9.17) yields

$$\mathcal{U}\big(S_\mu(\tau)\mathbf{y}\big) \;=\; U(\Phi y^0) + \sum_{s(\alpha)\sim 0} \frac{e^{is(\alpha)\cdot\mu\tau}}{m!}\, U^{(m)}(\Phi y^0)(\Phi\mathbf{y})^\alpha \,. \tag{9.23}$$

If $\mu \perp \mathcal{M}$, then the relation $s(\alpha)\sim 0$ implies $s(\alpha)\cdot\mu = 0$, and hence the expression (9.23) is independent of τ. It therefore follows that

$$0 = \frac{d}{d\tau}\Big|_{\tau=0} \mathcal{U}(S_\mu(\tau)\mathbf{y}) = \sum_{k\in\mathcal{N}} i(k\cdot\mu)\,\big(y^k\big)^T \nabla_k \mathcal{U}(\mathbf{y})$$

for all vectors $\mathbf{y} = (y^k)_{k\in\mathcal{N}}$. If μ is not orthogonal to $\mathcal{M}$, some terms in the sum of (9.23) depend on τ. However, for these terms with $s(\alpha)\in\mathcal{M}$ and $s(\alpha)\cdot\mu \neq 0$ we have $|s(\alpha)| \geq M = \min\{|k| \,:\, 0\neq k\in\mathcal{M}\}$ and if $\mu\perp\mathcal{M}_N$, then $|s(\alpha)|\geq N+1$. The bounds (5.13) then yield

$$\sum_{k\in\mathcal{N}} i(k\cdot\mu)\,\big(y^k\big)^T \nabla_k \mathcal{U}(\mathbf{y}) = \begin{cases} \mathcal{O}(\varepsilon^M) & \text{for arbitrary } \mu \\ \mathcal{O}(\varepsilon^{N+1}) & \text{for } \mu\perp\mathcal{M}_N \end{cases} \tag{9.24}$$

for the vector $\mathbf{y} = \mathbf{y}(t)$ as given by Theorem 9.2. Multiplying the relation (9.18) by $\frac{i}{\varepsilon}(-k\cdot\mu)\,\big(y^{-k}\big)^T$ and summing over $k\in\mathcal{N}$, we obtain with (9.24) that

$$-\,\frac{i}{\varepsilon}\sum_{k\in\mathcal{N}} (k\cdot\mu)(y^{-k})^T \Psi^{-1}\Phi\, h^{-2} L(hD) y^k \;=\; \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}).$$

The $\mathcal{O}(\varepsilon^{M-1})$ term is not present for $\mu\perp\mathcal{M}_N$. Written in the z variables, this becomes

$$-\,\frac{i}{\varepsilon}\sum_{k\in\mathcal{N}} (k\cdot\mu)(z^{-k})^T \Psi^{-1}\Phi\, h^{-2} L(hD + ihk\cdot\omega) z^k \;=\; \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}). \tag{9.25}$$

As in (9.19), the left-hand expression turns out to be the time derivative of a function $\mathcal{I}^*_\mu[\mathbf{z}](t)$ which depends on the values at t of the function $\mathbf{z}$ and its first N derivatives:

$$\frac{d}{dt}\mathcal{I}^*_\mu[\mathbf{z}](t) \;=\; \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}).$$

Together with Theorem 9.2 this yields the following.

Lemma 9.4. *Under the assumptions of Theorem 9.2, the modulation functions* $\mathbf{z}$ *satisfy*

$$\mathcal{I}^*_\mu[\mathbf{z}](t) \;=\; \mathcal{I}^*_\mu[\mathbf{z}](0) \;+\; \mathcal{O}(th^N) + \mathcal{O}(t\varepsilon^{M-1}) \tag{9.26}$$

for all $\mu\in\mathbb{R}^\ell$ *and for* $0\leq t\leq T$. *They satisfy*

$$\mathcal{I}^*_\mu[\mathbf{z}](t) \;=\; \mathcal{I}^*_\mu[\mathbf{z}](0) \;+\; \mathcal{O}(th^N) \tag{9.27}$$

for $\mu \perp \mathcal{M}_N$ *and* $0 \le t \le T$. *Moreover, at* $t = nh$,

$$\mathcal{I}_\mu^*[\mathbf{z}](t) \;=\; I_\mu^*(x_n, \dot{x}_n) + \mathcal{O}(\varepsilon), \tag{9.28}$$

where, again with $\sigma(\xi) = \mathrm{sinc}(\xi)\phi(\xi)/\psi(\xi)$,

$$I_\mu^*(x, \dot{x}) \;=\; \sum_{j=1}^{\ell} \sigma(\xi_j)\, \frac{\mu_j}{\lambda_j}\, I_j(x, \dot{x}). \tag{9.29}$$

XIII.9.4 Long-Time Near-Conservation of Total and Oscillatory Energies

With the proof of Theorem XIII.7.1, the above two lemmas yield the following results from Cohen, Hairer & Lubich (2004).

Theorem 9.5. *Under conditions (9.9)–(9.11), the numerical solution obtained by method (2.6) with (2.11) satisfies, for* H^* *and* I_μ^* *defined by (9.22) and (9.29),*

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= H^*(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I_\mu^*(x_n, \dot{x}_n) &= I_\mu^*(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \qquad \textit{for} \quad 0 \le nh \le h^{-N+1}$$

for $\mu \in \mathbb{R}^\ell$ *with* $\mu \perp \mathcal{M}_N = \{k \in \mathcal{M} : |k| \le N\}$. *The constants symbolized by* $\mathcal{O}$ *are independent of* n, h, ε, λ_j *satisfying the above conditions, but depend on* N *and the constants in the conditions.*

Since $\mu = \lambda$ is always orthogonal to $\mathcal{M}$ and to $\mathcal{M}_N$, the relation

$$K(x, \dot{x}) = H^*(x, \dot{x}) - I_\lambda^*(x, \dot{x})$$

for the smooth energy (9.7) implies

$$K(x_n, \dot{x}_n) = K(x_0, \dot{x}_0) + \mathcal{O}(h) \qquad \text{for} \quad 0 \le nh \le h^{-N+1}. \tag{9.30}$$

For $\sigma(\xi) = 1$ (or equivalently $\psi(\xi) = \mathrm{sinc}(\xi)\phi(\xi)$) the modified energies H^* and I_μ^* are identical to the original energies H and I_μ of (9.2) and (9.5). The condition $\psi(\xi) = \mathrm{sinc}(\xi)\phi(\xi)$ is known to be equivalent to the symplecticity of the one-step method $(x_n, \dot{x}_n) \mapsto (x_{n+1}, \dot{x}_{n+1})$, but its appearance in the above theorem is caused by a different mechanism which is not in any obvious way related to symplecticity. Without this condition we still have the following result, which also considers the long-time near-conservation of the individual oscillatory energies I_j for $j = 1, \ldots, \ell$.

Theorem 9.6. *Under conditions (9.9)–(9.11), the numerical solution obtained by method (2.6) with (2.11) satisfies*

$$\begin{aligned} H(x_n, \dot{x}_n) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I_j(x_n, \dot{x}_n) &= I_j(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \qquad \textit{for} \quad 0 \le nh \le h \cdot \min(\varepsilon^{-M+1}, h^{-N})$$

for $j = 1, \dots, \ell$, *with* $M = \min\{|k| : 0 \neq k \in \mathcal{M}\}$. *The constants symbolized by* $\mathcal{O}$ *are independent of* n, h, ε, λ_j *satisfying the above conditions, but depend on* N *and the constants in the conditions.*

For the non-resonant case $\mathcal{M} = \{0\}$ we have $M = \infty$ and hence the length of the interval with energy conservation is only restricted by (9.10). Notice that always $M \geq 3$, and $M = 3$ only in the case of a 1:2 resonance among the λ_j. For a 1:3 resonance we have $M = 4$ and in all other cases $M \geq 5$.

Explanation of the Numerical Experiment of Sect. XIII.9.1. All numerical methods in Figs. 9.2–9.4 satisfy the conditions of Theorems 9.6 and 9.5 for the step sizes considered.

In Fig. 9.2 we have the (symplectic) method (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \mathrm{sinc}(\xi)$, which has $\sigma(\xi) = 1$, so that H and H^*, and I_μ and I_μ^* coincide. For all step sizes, the oscillatory energy I_2 corresponding to the non-resonant frequency $\sqrt{2}\omega$ and the sum $I_1 + I_3$ are well conserved on long time intervals, in accordance with Theorem 9.5. The individual energies I_1 and I_3 corresponding to the resonant frequencies $\omega = 1/\varepsilon$ and $2/\varepsilon$ are not preserved on the time scale considered here, cf. Fig. 9.1. In fact, Theorem 9.6 here yields only a time scale $\mathcal{O}(h\varepsilon^{-2})$.

In Fig. 9.3 we use the method with $\phi(\xi) = 1$ and $\psi(\xi) = \mathrm{sinc}^2(\xi/2)$, for which $\sigma(\xi)$ is not identical to 1, and hence H and H^*, and I_μ and I_μ^* do not coincide. The oscillatory energy $I_2 = \sigma_2^{-1} I_\mu^*$ with $\mu = (0, 1, 0) \perp \mathcal{M}$, which corresponds to the non-resonant frequency $\sqrt{2}\omega$, is approximately conserved over long times. Since Theorem 9.5 only states that the *modified* energies are well preserved, it is not surprising that neither $I_1 + I_3$ nor the original total energy H (not shown in the figure) are conserved. The modified energies H^* and $\sigma_1 I_1 + \sigma_3 I_3$ (not shown) are indeed well conserved, and so is the smooth energy K, in agreement with (9.30).

Figure 9.4 shows the result for the (symplectic) method with $\phi(\xi) = \mathrm{sinc}(\xi)$ and $\psi(\xi) = \mathrm{sinc}(\xi)\phi(\xi)$. Since $\sigma(\xi) = 1$, the oscillatory energy I_2 for $\sqrt{2}\omega$ and also $I_1 + I_3$ are well conserved, in agreement with Theorem 9.5. However, the energy exchange between the resonant frequencies is not correctly reproduced. This behaviour is not explained by Theorems 9.5 and 9.6, but it corresponds to the analysis in Sect. XIII.4.2 which, for the single-frequency case, explains the incorrect energy exchange of methods that do not satisfy $\psi(\xi)\phi(\xi) = \mathrm{sinc}(\xi)$ (and thus, of all symplectic methods (2.7)–(2.10), with the exception of the above method with $\phi(\xi) = 1$ and $\psi(\xi) = \mathrm{sinc}(\xi)$). That analysis could be extended to the multi-frequency case considered here.

We remark that the techniques of Sects. XIII.9.2 and XIII.9.3 can also be used to study the energy error of the Störmer–Verlet method, as in Sect. XIII.8; see Theorem 5.1 in Cohen, Hairer & Lubich (2004). The modulated Fourier expansion of the exact solution yields results on the near-preservation of the oscillatory energies along a bounded exact solution: under the energy bound (9.9) and the non-resonance condition

$$|k \cdot \lambda| \geq c\sqrt{\varepsilon} \quad \text{for } k \in \mathbb{Z}^\ell \setminus \mathcal{M} \text{ with } |k| \leq N \tag{9.31}$$

we have (see Theorem 6.1 in Cohen, Hairer & Lubich 2004)

$$I_\mu\big(x(t),\dot x(t)\big) \;=\; I_\mu\big(x(0),\dot x(0)\big) + \mathcal{O}(\varepsilon) \quad \text{for } 0 \le t \le \varepsilon^{-N+1} \tag{9.32}$$

for $\mu \in \mathbb{R}^\ell$ with $\mu \perp \mathcal{M}_N = \{k \in \mathcal{M} : |k| \le N\}$. We further have

$$I_j\big(x(t),\dot x(t)\big) \;=\; I_j\big(x(0),\dot x(0)\big) + \mathcal{O}(\varepsilon) \quad \text{for } 0 \le t \le \varepsilon\cdot\min(\varepsilon^{-M+1},\varepsilon^{-N}) \tag{9.33}$$

for $j = 1,\ldots,\ell$, with $M = \min\{|k| : 0 \ne k \in \mathcal{M}\}$.

XIII.10 Systems with Non-Constant Mass Matrix

The high frequencies of the linearized differential equation remain constant up to small deviations for mechanical systems with a Hamiltonian of the form

$$H(p,q) = \frac{1}{2}p_0^T M_0(q)^{-1}p_0 + \frac{1}{2}p_1^T M_1^{-1}p_1 + \frac{1}{2}p^T R(q)p + \frac{1}{2\varepsilon^2}q_1^T A_1 q_1 + U(q) \tag{10.1}$$

with a symmetric positive definite matrix $M_0(q)$, constant symmetric positive definite matrices M_1 and A_1, a symmetric matrix $R(q)$ with

$$R(q_0, 0) = 0,$$

and a potential $U(q)$. All the functions are assumed to depend smoothly on q. Bounded energy then requires $q_1 = \mathcal{O}(\varepsilon)$, so that $p^T R(q)p = \mathcal{O}(\varepsilon)$, but the derivative of this term with respect to q_1 is $\mathcal{O}(1)$.

As in (9.1), we may assume, after an appropriate canonical linear transformation based on a Cholesky decomposition of the mass matrix and a diagonalization of the resulting stiffness matrix, that the Hamiltonian is of the form

$$H(p,q) = \frac{1}{2}p_0^T M_0(q)^{-1}p_0 + \frac{1}{2}\sum_{j=1}^{\ell}\Big(\|p_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2}\,\|q_j\|^2\Big) + \frac{1}{2}p^T R(q)p + U(q) \tag{10.2}$$

with distinct, constant $\lambda_j \ge 1$.

The necessity for such a generalization results from the fact that oscillatory mechanical systems with near-constant frequencies in 2 or 3 space dimensions typically cannot be put in the form (9.1), but in the more general form (10.1) or (10.2).

Example 10.1 (Stiff Spring Pendulum). The motion of a mass point (of mass 1) hanging on a massless stiff spring (with spring constant $1/\varepsilon^2$) is described in polar coordinates $x_1 = r\sin\varphi$, $x_2 = -r\cos\varphi$ by the Lagrangian with kinetic energy $T = \frac{1}{2}(\dot x_1^2 + \dot x_2^2) = \frac{1}{2}(\dot r^2 + r^2\dot\varphi^2)$ and potential energy $U = \frac{1}{2\varepsilon^2}\,(r-1)^2 - r\cos\varphi$. With the coordinates $q_0 = \varphi$, $q_1 = r-1$ and the conjugate momenta $p_i = \partial T/\partial q_i$ this gives the Hamiltonian

$$H(p,q) = \frac{1}{2}\big((1+q_1)^{-2}p_0^2 + p_1^2\big) + \frac{1}{2\varepsilon^2}\,q_1^2 - (1+q_1)\cos q_0,$$

which is of the form (10.2).

Numerical methods for systems (10.2) are studied by Cohen (2005). He splits the small term $\frac{1}{2}p^T R(q)p$ from the principal terms of the Hamiltonian and proposes the following method, where

$$K(p_0, q) = \frac{1}{2} p_0^T M_0(q)^{-1} p_0 + U(q).$$

Algorithm 10.2. *1. A half-step with the symplectic Euler method applied to the system with Hamiltonian $\frac{1}{2}p^T R(q)p$ gives*

$$\begin{aligned} \widehat{p}^n &= p^n - \frac{h}{2} \nabla_q \Big(\frac{1}{2} (\widehat{p}^n)^T R(q^n) \widehat{p}^n \Big) \\ \widehat{q}^n &= q^n + \frac{h}{2} R(q^n) \widehat{p}^n. \end{aligned} \tag{10.3}$$

2. Treating the oscillatory components of the variables p and q with a trigonometric method (2.7)–(2.8) and the slow components with the Störmer-Verlet scheme yields (for $j = 1, \ldots, \ell$ and with $\omega_j = \lambda_j/\varepsilon$ and $\xi_j = h\omega_j$)

$$\begin{aligned} p_0^{n+1/2} &= \widehat{p}_0^n - \frac{h}{2} \nabla_{q_0} K(p_0^{n+1/2}, \Phi \widehat{q}^n) \\ \widehat{q}_0^{n+1} &= \widehat{q}_0^n + \frac{h}{2} \Big(\nabla_{p_0} K(p_0^{n+1/2}, \Phi \widehat{q}^n) + \nabla_{p_0} K(p_0^{n+1/2}, \Phi \widehat{q}^{n+1}) \Big) \\ \widehat{q}_j^{n+1} &= \cos(\xi_j) \widehat{q}_j^n + \omega_j^{-1} \sin(\xi_j) \widehat{p}_j^n - \frac{h^2}{2} \psi(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \widehat{q}^n) \\ \widehat{p}_j^{n+1} &= -\omega_j \sin(\xi_j) \widehat{q}_j^n + \cos(\xi_j) \widehat{p}_j^n - \frac{h}{2} \Big(\psi_0(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \widehat{q}^n) \\ &\quad + \psi_1(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \widehat{q}^{n+1}) \Big), \\ \widehat{p}_0^{n+1} &= p_0^{n+1/2} - \frac{h}{2} \nabla_{q_0} K(p_0^{n+1/2}, \Phi \widehat{q}^{n+1}) \end{aligned} \tag{10.4}$$

where $\Phi = \phi(h\Omega)$ with $\Omega = \mathrm{diag}(\omega_j I)$.

3. A half-step with the adjoint symplectic Euler method applied to the system with Hamiltonian $\frac{1}{2}p^T R(q)p$ gives

$$\begin{aligned} p^{n+1} &= \widehat{p}^{n+1} - \frac{h}{2} \nabla_q \Big(\frac{1}{2} (\widehat{p}^{n+1})^T R(q^{n+1}) \widehat{p}^{n+1} \Big) \\ q^{n+1} &= \widehat{q}^{n+1} + \frac{h}{2} R(q^{n+1}) \widehat{p}^{n+1}. \end{aligned} \tag{10.5}$$

The filter functions $\psi, \psi_0, \psi_1, \phi$ are again real-valued functions with $\psi(0) = \hat{\psi}(0) = \tilde{\psi}(0) = \phi(0) = 1$ that satisfy (2.9). The method is still symplectic if and only if (2.10) holds. Note that Step 2. of the algorithm is explicit if $M_0(q)$ does not depend on q_0.

Cohen (2004, 2005) studies the modulated Fourier expansion of this method and shows that the long-time near-conservation of total and oscillatory energies as given by Theorem 9.6 remains valid also in this more general situation.

Example 10.3 (Triatomic Molecule). The motion of a near-rigid triatomic molecule is described by a Hamiltonian system with a Hamiltonian (10.2). For simplicity we fix the position of the central atom. We then have two stiff-spring pendulums strongly coupled by another spring. With angles and distances as shown in Fig. 10.1, we use the position coordinates $\varphi_1, q_1 = r_1 - 1, \varphi_2, q_2 = r_2 - 1$ with the conjugate momenta π_1, p_1, π_2, p_2, respectively. The Hamiltonian then reads

$$\begin{aligned} H(\pi, p, \varphi, q) &= \frac{1}{2}\Big((1+q_1)^{-2}\pi_1^2 + p_1^2 + (1+q_2)^{-2}\pi_2^2 + p_2^2\Big) \\ &\quad + \frac{1}{2\varepsilon^2}\Big(q_1^2 + q_2^2 + \frac{\alpha^2}{2}(\varphi_2 - \varphi_1)^2\Big) + U(\varphi, q) \end{aligned} \tag{10.6}$$

with a spring constant $\frac{1}{2}\alpha^2/\varepsilon^2$ for connecting the two pendulums and an external potential U. With the canonical change of variables

$$\begin{pmatrix} q_3 \\ q_0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \qquad \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} p_3 \\ p_0 \end{pmatrix},$$

the Hamiltonian takes the form (10.2):

$$\begin{aligned} H(p, q) &= \frac{1}{2}(p_0^2 + p_1^2 + p_2^2 + p_3^2) + \frac{1}{2\varepsilon^2}(q_1^2 + q_2^2 + \alpha^2 q_3^2) \\ &\quad + p^T R(q) p + \widehat{U}(q) \end{aligned} \tag{10.7}$$

with

$$p^T R(q) p = -\frac{1}{4}\frac{2q_2 + q_2^2}{(1+q_2)^2}(p_0 - p_3)^2 - \frac{1}{4}\frac{2q_1 + q_1^2}{(1+q_1)^2}(p_0 + p_3)^2$$

and $\widehat{U}(q) = U(\varphi_1, \varphi_2, q_1, q_2)$.

For the water molecule the ratio between the frequencies of the bond angle and the bond lengths is $\alpha \approx 0.2$, according to some popular models. In our numerical experiments, we observed good conservation of all the oscillatory energies and the total energy. More interesting phenomena occur in a near-resonance situation. We consider $\alpha = 0.49$ and $\varepsilon = 0.01$, no exterior potential ($U = 0$), and initial values $q(0) = (0, \varepsilon/2, \varepsilon, \alpha/\varepsilon)$ and $p(0) = (1.1, 0.2, -0.8, 1.3)$. In Fig. 10.2 we apply the method of Algorithm 10.2 with step sizes $h = 0.5\,\varepsilon$ and $h = 2\varepsilon$ and obtain

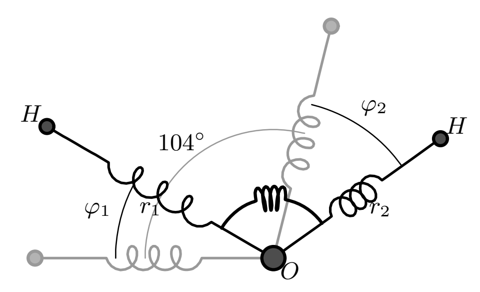


Fig. 10.1. Water molecule and reference configuration as gray shadow

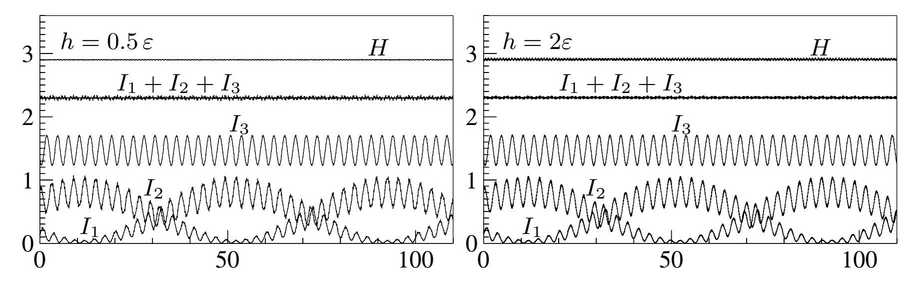


Fig. 10.2. Oscillatory energies and total energy for the method of Algorithm 10.2

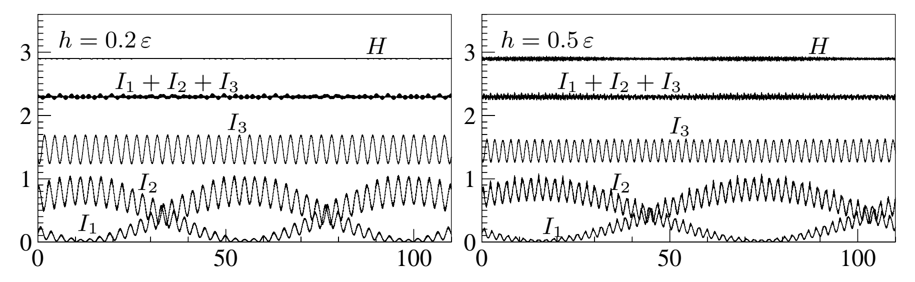


Fig. 10.3. Oscillatory energies and total energy for the Störmer–Verlet method

numerical results that agree very well with a solution obtained with very small step sizes. For comparison we show in Fig. 10.3 the results of the Störmer–Verlet method with step sizes $h = 0.2\,\varepsilon$ and $h = 0.5\,\varepsilon$, for which the energy exchange is not correct. For the reason explained in Sect. VI.3, (3.2)–(3.3), both methods are fully explicit for this problem.

XIII.11 Exercises

1. Show that the impulse method (with exact solution of the fast system) reduces to Deuflhard's method in the case of a quadratic potential $W(q) = \frac{1}{2} q^T A q$.
2. Show that a method (2.7)–(2.8) satisfying (2.9) is symplectic if and only if

$$\psi(\xi) = \operatorname{sinc}(\xi)\, \phi(\xi) \qquad \text{for} \quad \xi = h\omega .$$

3. The change of coordinates $x_n = \chi(h\Omega) z_n$ transforms (2.7)–(2.8) into a method of identical form with $\phi, \psi, \psi_0, \psi_1$ replaced by $\chi\phi, \chi^{-1}\psi, \chi^{-1}\psi_0, \chi^{-1}\psi_1$. Prove that, for $h\omega$ satisfying $\operatorname{sinc}(h\omega)\phi(h\omega)/\psi(h\omega) > 0$, it is possible to find $\chi(h\omega)$ such that the transformed method is symplectic.
4. Prove that for infinitely differentiable functions $g(t)$ the solution of $\ddot{x} + \omega^2 x = g(t)$ can be written as

$$x(t) \;=\; y(t) + \cos(\omega t)\, u(t) + \sin(\omega t)\, v(t),$$

where $y(t)$, $u(t)$, $v(t)$ are given by asymptotic expansions in powers of ω^{-1}.
Hint. Use the variation-of-constants formula and apply repeated partial integration.

5. Show that the recurrence relation $e_{n+1} - 2\cos(h\Omega)\, e_n + e_{n-1} = b_n$ has the solution

$$e_{n+1} = -W_{n-1}\, e_0 + W_n\, e_1 + \sum_{j=1}^{n} W_{n-j}\, b_j$$

with $W_n = \sin(h\Omega)^{-1} \sin\big((n+1)h\Omega\big)$ (or the appropriate limit when $\sin(h\Omega)$ is not invertible).

6. Consider a Hamiltonian $H(p_R, p_I, q_R, q_I)$ and let

$$\mathcal{H}(p,q) \;=\; 2\, H(p_R, p_I, q_R, q_I)$$

for $p = p_R + ip_I$, $q = q_R + iq_I$. Prove that in the new variables p, q the Hamiltonian system becomes

$$\dot p = -\frac{\partial \mathcal{H}}{\partial \overline{q}}(p,q), \qquad \dot q = \frac{\partial \mathcal{H}}{\partial \overline{p}}(p,q).$$

7. Prove the following refinement of Theorem 6.3: along the solution $x(t)$ of (2.1), the modified oscillatory energy $J(x,\dot x) = I(x,\dot x) - x_1^T g_1(x)$ satisfies

$$J(x(t),\dot x(t)) = J(x(0),\dot x(0)) + \mathcal{O}(\omega^{-2}) + \mathcal{O}(t\omega^{-N})\ .$$

8. Define $\widehat H(x,\dot x) = H(x,\dot x) - \rho x_1^T g_1(x)$, $\widehat J(x,\dot x) = J(x,\dot x) - \rho x_1^T g_1(x)$ with $J(x,\dot x)$ of the previous exercise and with

$$\rho = \frac{\psi(h\omega)}{\mathrm{sinc}^2(\frac12 h\omega)} - 1.$$

In the situation of Theorem 7.1, show that

$$\begin{array}{rcl} \widehat H(x_n,\dot x_n) &=& \widehat H(x_0,\dot x_0) + \mathcal{O}(h^2) \\ \widehat J(x_n,\dot x_n) &=& \widehat J(x_0,\dot x_0) + \mathcal{O}(h^2) \end{array} \qquad \text{for} \quad 0 \le nh \le h^{-N+1}.$$

Notice that the total energy $H(x_n,\dot x_n)$ and the modified oscillatory energy $J(x_n,\dot x_n)$ are conserved up to $\mathcal{O}(h^2)$ if $\rho = 0$, i.e., if $\psi(\xi) = \mathrm{sinc}^2(\frac12\xi)$. This explains the excellent energy conservation of methods (A) and (D) in Figure 2.5 away from resonances.

9. Generalizing the analysis of Sect. XIII.8, study the energy behaviour of the impulse or averaged-force multiple time-stepping method of Sect. VIII.4 with a fixed number N of Störmer–Verlet substeps per step, when the method is applied to the model problem with $h\omega$ bounded away from zero.

Chapter XIV. Oscillatory Differential Equations with Varying High Frequencies

New aspects come into play when the high frequencies in an oscillatory system and their associated eigenspaces do not remain nearly constant, as in the previous chapter, but change with time or depend on the solution. We begin by studying linear differential equations with a time-dependent skew-hermitian matrix and then turn to nonlinear oscillatory mechanical systems with time- or solution-dependent frequencies. Our analysis uses canonical coordinate transforms that separate slow and fast motions and relate the fast oscillations to the skew-hermitian linear case. For the numerical treatment we consider suitably constructed long-time-step methods ("adiabatic integrators") and multiple time-stepping methods.

XIV.1 Linear Systems with Time-Dependent Skew-Hermitian Matrix

We consider first-order linear differential equations with a skew-hermitian matrix that changes slowly compared to the rapid oscillations in the solution, a problem that has attracted much attention in quantum mechanics. We present a suitable class of numerical methods, termed adiabatic integrators, which can take time steps that are substantially larger than the almost-periods of the oscillations.

XIV.1.1 Adiabatic Transformation and Adiabatic Invariants

> It comes from the greek $\alpha\delta\iota\alpha\beta\alpha\tau\iota\chi o\varsigma$, "which cannot be crossed".
>
> ... we arrive by analogy to the "adiabatic principle" used in Quantum and then Classical Mechanics. It is based upon the fact that the harmonic oscillator (and other simple dynamical systems as it was found later) submitted to slow variations of its parameters modifies its energy but keeps its action (energy divided by frequency) constant.
>
> As we can see, the path from the word "adiabatic" used in thermodynamics to the above "adiabatic principle" is tortuous and our greek colleagues are certainly puzzled by sentences such as "the changes in the adiabatic invariant due to [...] crossing" which we shall use later.
>
> (J. Henrard 1993)

We consider the linear differential equation

$$\dot{y}(t) = \frac{1}{\varepsilon} Z(t)\, y(t), \tag{1.1}$$

where $Z(t)$ is a real skew-symmetric (or complex skew-hermitian) matrix-valued function with time derivatives bounded independently of the small parameter ε. In quantum dynamics such equations arise with $Z(t) = -iH(t)$, where the real symmetric (or hermitian) matrix $H(t)$ represents the quantum Hamiltonian operator in a discrete-level Schrödinger equation. We will also encounter real equations of this type in the treatment of oscillatory classical mechanical systems with time-dependent frequencies. Solutions oscillate with almost-periods $\sim \varepsilon$, while the system matrix changes on a slower time scale ~ 1.

Transforming the Problem. We begin by looking for a time-dependent linear transformation

$$\eta(t) = T_\varepsilon(t) y(t), \tag{1.2}$$

taking the system to the form

$$\dot{\eta}(t) = S_\varepsilon(t)\, \eta(t) \qquad \text{with} \qquad S_\varepsilon = \dot{T}_\varepsilon T_\varepsilon^{-1} + \frac{1}{\varepsilon} T_\varepsilon Z T_\varepsilon^{-1}, \tag{1.3}$$

which is chosen such that $S_\varepsilon(t)$ is of smaller norm than the matrix $\frac{1}{\varepsilon} Z(t)$ of (1.1).

Remark 1.1. A first idea is to freeze $Z(t) \approx Z_*$ over a time step and to choose the transformation

$$T_\varepsilon(t) = \exp\Big(-\frac{t}{\varepsilon} Z_*\Big) \quad \text{yielding} \quad S_\varepsilon(t) = \frac{1}{\varepsilon} \exp\Big(-\frac{t}{\varepsilon} Z_*\Big) \big(Z(t) - Z_*\big) \exp\Big(\frac{t}{\varepsilon} Z_*\Big).$$

This matrix function $S_\varepsilon(t)$ is highly oscillatory and bounded in norm by $\mathcal{O}(h/\varepsilon)$ for $|t - t_0| \le h$, if $Z_* = Z(t_0 + h/2)$. Numerical integrators based on this transformation are given by Lawson (1967) and more recently by Hochbruck & Lubich (1999b), Iserles (2002, 2004), and Degani & Schiff (2003). Reasonable accuracy still requires step sizes $h = \mathcal{O}(\varepsilon)$ in general; see also Exercise 3. In the above papers this transfomation has, however, been put to good use in situations where the time derivatives of the matrix in the differential equation have much smaller norm than the matrix itself.

Adiabatic Transformation. In order to obtain a differential equation (1.3) with a uniformly bounded matrix $S_\varepsilon(t)$ we diagonalize

$$Z(t) = U(t)\, i\Lambda(t)\, U(t)^*$$

with a real diagonal matrix $\Lambda(t) = \operatorname{diag}(\lambda_j(t))$ and a unitary matrix $U(t) = (u_1(t), \ldots, u_n(t))$ of eigenvectors depending smoothly on t (possibly except where eigenvalues cross). We define $\eta(t)$ by the unitary *adiabatic transformation*

$$\eta(t) = \exp\Big(-\frac{i}{\varepsilon}\Phi(t)\Big) U(t)^* y(t) \quad \text{with} \quad \Phi(t) = \operatorname{diag}(\phi_j(t)) = \int_0^t \Lambda(s)\, ds, \tag{1.4}$$

which represents the solution in a rotating frame of eigenvectors. Each component of $\eta(t)$ is a coefficient in the eigenbasis representation of $y(t)$ rotated in the complex plane by the negative phase. Such transformations have been in use in quantum mechanics since the work of Born & Fock (1928) on adiabatic invariants in Schrödinger equations, as discussed in the next paragraph. The transformation (1.4) yields a differential equation where the ε-independent skew-hermitian matrix

$$W(t) = \dot{U}(t)^* U(t)$$

is framed by oscillatory diagonal matrices:

$$\dot{\eta}(t) = \exp\Bigl(-\frac{i}{\varepsilon}\Phi(t)\Bigr)\, W(t)\, \exp\Bigl(\frac{i}{\varepsilon}\Phi(t)\Bigr)\, \eta(t). \tag{1.5}$$

Numerical integrators for (1.1) based on the transformation to the differential equation (1.5) with bounded, though highly oscillatory right-hand side, are given by Jahnke & Lubich (2003) and Jahnke (2004a); see Sect. XIV.1.2.

Adiabatic Invariants. Possibly after a time-dependent rephasing of the eigenvectors, $u_k(t) \to e^{i\alpha_k(t)} u_k(t)$, we can assume that $\dot{u}_k(t)$ is orthogonal to $u_k(t)$ for all t. (This is automatically satisfied if $U(t)$ is a real orthogonal matrix, as is the case for $Z(t) = -iH(t)$ with a real symmetric matrix $H(t)$.) We then have the matrix $W = (w_{jk}) = (\dot{u}_j^* u_k)$ with zero diagonal.

After integration of both sides of the differential equation (1.5) from 0 to t, partial integration of the terms on the right-hand side yields for $j \neq k$ (terms for $j = k$ do not appear since $w_{jj} = 0$)

$$\begin{aligned} &\int_0^t \exp\Bigl(-\frac{i}{\varepsilon}\bigl(\phi_j(s) - \phi_k(s)\bigr)\Bigr)\, w_{jk}(s)\, \eta_k(s)\, ds \\ &= \; i\varepsilon\, \exp\Bigl(-\frac{i}{\varepsilon}\bigl(\phi_j(s) - \phi_k(s)\bigr)\Bigr)\, \frac{w_{jk}(s)\, \eta_k(s)}{\lambda_j(s) - \lambda_k(s)}\, \Bigg|_0^t \\ &\quad - i\varepsilon \int_0^t \exp\Bigl(-\frac{i}{\varepsilon}\bigl(\phi_j(s) - \phi_k(s)\bigr)\Bigr)\, \frac{d}{ds}\frac{w_{jk}(s)\, \eta_k(s)}{\lambda_j(s) - \lambda_k(s)}\, ds\,. \end{aligned} \tag{1.6}$$

At this point, suppose that the eigenvalues $\lambda_j(t)$ are, for all t, separated from each other by a positive distance δ independent of ε:

$$|\lambda_j(t) - \lambda_k(t)| \geq \delta \qquad \text{for all } \; j \neq k. \tag{1.7}$$

Then the reciprocals of their differences and the coupling matrix $W(t)$ are bounded independently of ε, as are their derivatives. Together with the boundedness of $\dot{\eta}$ as implied by (1.5), this shows

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \leq \text{Const.} \tag{1.8}$$

This result is a version of the quantum-adiabatic theorem of Born & Fock (1928) which states that the actions $|\eta_j|^2$ (the energy in the jth state, $\langle \eta_j u_j, H\eta_j u_j \rangle =$

$\lambda_j|\eta_j|^2$, divided by the frequency λ_j) remain approximately constant for times $t = \mathcal{O}(1)$. Such functions $I(y,t)$ that satisfy $I(y(t),t) = I(y(0),0) + \mathcal{O}(\varepsilon)$ for $t = \mathcal{O}(1)$ along every $\mathcal{O}(1)$-bounded solution $y(t)$ of the differential equation, are called *adiabatic invariants.*

Super-Adiabatic Transformations. Adiabatic invariants are obtained over longer time scales by refining the transformation; see Lenard (1959) and Garrido (1964). Here we show that the transformation matrix T_ε of (1.2) can be constructed such that the matrix S_ε in the transformed differential equation (1.3) is of size $\mathcal{O}(\varepsilon^N)$. Let us make the ansatz of a unitary transformation matrix

$$T_\varepsilon^{(N)} = \exp\Big(-\frac{i}{\varepsilon}\Phi\Big)\exp(-i\Phi_1)\exp(\varepsilon X_1)\,\dots\,\exp(-i\varepsilon^{N-1}\Phi_N)\exp(\varepsilon^N X_N)\,U^*$$

with real diagonal matrices $\Phi_n(t)$ and complex skew-hermitian matrices $X_n(t)$. We find that $S_\varepsilon = \frac{1}{\varepsilon}T_\varepsilon Z T_\varepsilon^* + \dot T_\varepsilon T_\varepsilon^*$ is $\mathcal{O}(\varepsilon)$ if and only if X_1 and $\Lambda_1 := \dot\Phi_1$ satisfy

$$\frac{1}{\varepsilon}\Big(\exp(\varepsilon X_1)\, i\Lambda\, \exp(-\varepsilon X_1) - i\Lambda\Big) - i\Lambda_1 + W = \mathcal{O}(\varepsilon),$$

or equivalently, if X_1 and Λ_1 solve the commutator equation

$$[i\Lambda, X_1] + i\Lambda_1 = W.$$

This is solved by setting $i\Lambda_1$ equal to the diagonal of W and determining the off-diagonal entries $x_{jk}^{(1)}$ of X_1 from the scalar equations

$$i(\lambda_j - \lambda_k)\, x_{jk}^{(1)} = w_{jk}, \qquad j \neq k,$$

which can be done as long as the eigenvalues are separated. The diagonal of X_1 is set to zero. Since W is skew-hermitian, so is X_1. Similarly we obtain for higher powers of ε the equations

$$[i\Lambda, X_n] + i\Lambda_n = W_{n-1},$$

where the matrix W_{n-1} contains only previously constructed terms up to index $n-1$ and derivatives up to order n and is skew-hermitian because S_ε is skew-hermitian. In this way we obtain a unitary transformation such that

$$\eta^{(N)}(t) = T_\varepsilon^{(N)}(t)\, y(t) \qquad \text{satisfies} \qquad \dot\eta^{(N)} = \mathcal{O}(\varepsilon^N).$$

We remark that the above construction of $T_\varepsilon^{(N)}$ is analogous to transformations in Hamiltonian perturbation theory; cf. Sect. X.2.

The differential equation (1.1) thus has adiabatic invariants over times $\mathcal{O}(\varepsilon^{-N})$ for arbitrary $N \geq 1$, and in fact even over exponentially long time intervals $t = \mathcal{O}(e^{c/\varepsilon})$ if the functions have a bounded analytic extension to a complex strip, as is shown by Joye & Pfister (1993) and Nenciu (1993). The leading term in the exponentially small deviation of $|\eta_j^{(N)}(t)|^2$ in the optimally truncated super-adiabatic basis has been rigorously made explicit by Betz & Teufel (2005a, 2005b), proving a conjecture by Berry (1990).

Avoided Crossing of Eigenvalues and Non-Adiabatic Transitions. To illustrate the effects of a violation of the separation condition (1.7), we consider the generic two-dimensional example studied by Zener (1932), with the matrix

$$Z(t) = -i \begin{pmatrix} t & \delta \\ \delta & -t \end{pmatrix}, \tag{1.9}$$

which has the eigenvalues $\pm i\sqrt{t^2+\delta^2}$. The minimal distance of the eigenvalues is 2δ at $t=0$. For $\delta = \mathcal{O}(\sqrt{\varepsilon})$ the adiabatic invariance (1.8) is no longer valid, and η can undergo $\mathcal{O}(1)$ changes in an $\mathcal{O}(\delta)$ neighbourhood of $t=0$: a *non-adiabatic transition* in physical terminology. The changes in the adiabatic invariant due to the avoided crossing of eigenvalues are illustrated in Fig. 1.1 and can be explained as follows.

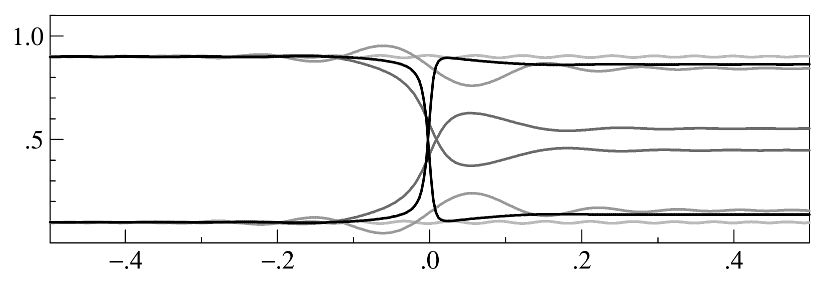


Fig. 1.1. Non-adiabatic transition: $|\eta_1(t)|^2$ and $|\eta_2(t)|^2$ as function of t for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ (increasing darkness)

Near the avoided crossing, a new time scale $\tau = t/\delta$ is appropriate. The decomposition $Z(t) = U(t) i\Lambda(t) U(t)^T$ of the matrix yields

$$\begin{aligned} U(t) \;=\; \widetilde{U}(\tau) \;&=\; \begin{pmatrix} \cos\alpha(\tau) & -\sin\alpha(\tau) \\ \sin\alpha(\tau) & \cos\alpha(\tau) \end{pmatrix}, \\ \Lambda(t)/\delta \;=\; \widetilde{\Lambda}(\tau) \;&=\; \begin{pmatrix} -\sqrt{\tau^2+1} & 0 \\ 0 & \sqrt{\tau^2+1} \end{pmatrix}, \end{aligned}$$

with $\alpha(\tau) = \frac{\pi}{4} - \frac{1}{2}\arctan(\tau)$. We introduce the rescaled matrices

$$\begin{aligned} \widetilde{\Phi}(\tau) \;&=\; \int_0^\tau \widetilde{\Lambda}(\sigma)\, d\sigma \;=\; \Phi(t)/\delta^2, \\ \widetilde{W}(\tau) \;&=\; \left(\frac{d}{d\tau}\widetilde{U}(\tau)^T\right)\widetilde{U}(\tau) \;=\; \frac{1}{2(\tau^2+1)}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \;=\; \delta\cdot W(t). \end{aligned}$$

Note that the entries of $W(t)$ have a sharp peak of height $(2\delta)^{-1}$ at $t=0$. The rescaled function $\widetilde{\eta}(\tau) = \eta(t)$ is a solution of the differential equation

$$\frac{d}{d\tau}\widetilde{\eta}(\tau) = \exp\left(-\frac{i\delta^2}{\varepsilon}\widetilde{\Phi}(\tau)\right)\widetilde{W}(\tau)\exp\left(\frac{i\delta^2}{\varepsilon}\widetilde{\Phi}(\tau)\right)\widetilde{\eta}(\tau).$$

For $\delta^2 \le \varepsilon$ and $|\tau| = |t/\delta| \le 1$, the matrix on the right-hand side is bounded of norm ~ 1 and has bounded derivatives with respect to τ. The function $\widetilde{\eta}(\tau)$ therefore changes its value by an amount of size $\mathcal{O}(1)$ in the interval $|\tau| \le 1$. We also note that any numerical integrator using piecewise polynomial approximations of $W(t)$ and hence of $\widetilde{W}(\tau)$ must take step sizes $\Delta\tau = h/\delta \ll 1$, i.e., $h \ll \delta$. On the other hand, the rescaling shows that the number of time steps needed to resolve the non-adiabatic transition up to a specified accuracy is independent of δ.

XIV.1.2 Adiabatic Integrators

We discuss symmetric long-time-step integrators for the rotating-frame differential equation (1.5) that describes skew-hermitian systems in adiabatic variables. The construction follows Jahnke & Lubich (2003) and Jahnke (2004a); see also Lorenz, Jahnke & Lubich (2005).

First-Order Integrators. We consider the differential equation (1.5) and integrate both sides from t_n to $t_{n+1} = t_n + h$:

$$\eta(t_{n+1}) = \eta(t_n) + \int_{t_n}^{t_{n+1}} \exp\Big(-\frac{i}{\varepsilon}\Phi(s)\Big)\, W(s)\, \exp\Big(\frac{i}{\varepsilon}\Phi(s)\Big)\, \eta(s)\, ds\,, \tag{1.10}$$

where $W(t)$ is an ε-independent matrix, continuously differentiable in t, and the real diagonal matrix of phases $\Phi(t)$ is given as the integral of $\Lambda(t) = \operatorname{diag}(\lambda_j(t))$. In the applications, $W(t)$ and $\Phi(t)$ are not given explicitly, but need to be computed using numerical differentiation and integration, respectively. For simplicity, we here ignore this approximation and consider W, Φ, Λ as given time-dependent functions.

Since η and W have bounded derivatives, the following averaged version of the implicit midpoint rule has a local error of $\mathcal{O}(h^2)$ uniformly in ε:[1]

$$\eta_{n+1} = \eta_n + \int_{t_n}^{t_{n+1}} \exp\Big(-\frac{i}{\varepsilon}\Phi(s)\Big)\, W(t_{n+1/2})\, \exp\Big(\frac{i}{\varepsilon}\Phi(s)\Big)\, ds\; \frac{1}{2}(\eta_{n+1} + \eta_n). \tag{1.11}$$

The problem then remains to compute the oscillatory integral. The integrand can be rewritten as

$$E(\Phi(s)) \bullet W(t_{n+1/2}),$$

where $\bullet$ denotes the entrywise product of matrices and

$$E(\Phi) = (e_{jk}) \qquad \text{with} \quad e_{jk} = \exp\Big(-\frac{i}{\varepsilon}(\phi_j - \phi_k)\Big).$$

With a linear phase approximation (of an error $\mathcal{O}(h^2)$)

$$\Phi(t_{n+1/2} + \theta h) \approx \Phi(t_{n+1/2}) + \theta h \Lambda(t_{n+1/2}),$$

[1] Because of the oscillatory integrals, the local error is not $\mathcal{O}(h^3)$ as might at first glance be expected for a symmetric method.

the integral is approximated by

$$h\, E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2})$$

where $\mathcal{I}(t)$ is the matrix of integrated exponentials with entries (we omit the argument t)

$$\mathcal{I}_{jk} = \int_{-1/2}^{1/2} \exp\Big(-\frac{i\theta h}{\varepsilon}(\lambda_j - \lambda_k)\Big) d\theta = \operatorname{sinc}\Big(\frac{h}{2\varepsilon}(\lambda_j - \lambda_k)\Big).$$

The error in the integral approximation comes solely from the linear phase approximation and is bounded by $\mathcal{O}\big(h \cdot \frac{h^2}{\varepsilon} \cdot \frac{\varepsilon}{h}\big) = \mathcal{O}(h^2)$ if the λ_j are separated, because then the integral $\mathcal{I}_{jk}$ is of size $\mathcal{O}\big(\frac{\varepsilon}{h}\big)$. We thus obtain the following *averaged implicit midpoint rule* with a local error of $\mathcal{O}(h^2)$ uniformly in ε:

$$\eta_{n+1} = \eta_n + h\Big(E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2})\Big)\, \frac{1}{2}(\eta_{n+1} + \eta_n). \quad (1.12)$$

An analogue of the explicit midpoint rule is similarly constructed, and from the Magnus series (IV.7.5) of the solution we obtain the following *averaged exponential midpoint rule*, again with an $\mathcal{O}(h^2)$ local error uniformly in ε:

$$\eta_{n+1} = \exp\Big(h\, E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2})\Big)\, \eta_n. \quad (1.13)$$

For skew-hermitian $W(t)$, also the matrix in (1.12) and (1.13) is skew-hermitian, and hence both of the above integrators preserve the Euclidean norm of η exactly. We summarize the local error bounds for these methods under conditions that include the case of an avoided crossing of eigenvalues.

Theorem 1.2 (Local Error). *Suppose that for* $t_0 \le t \le t_0 + h$ *and all* j, k,

$$|\lambda_j(t) - \lambda_k(t)| \ge \delta, \quad |\dot\lambda_j(t)| \le C_0, \quad \|W(t)\| \le \frac{C_1}{\delta}, \quad \|\dot W(t)\| \le \frac{C_2}{\delta^2}$$

with $\delta > 0$. *Then, the local error of methods (1.12) and (1.13) is bounded by*

$$\|\eta_1 - \eta(t_0 + h)\| \le C\, \frac{h^2}{\delta^2}\, \|\eta_0\|.$$

The constant C *is independent of* h, ε, δ.

Proof. The result is obtained with the arguments and approximation estimates given above, taking in addition account of the dependence on δ. □

The local error contains smooth, non-oscillatory components which accumulate to a global error $\eta_n - \eta(t_n) = \mathcal{O}(h)$ on bounded intervals if the eigenvalues remain well separated. Using that in this case η is constant up to $\mathcal{O}(\varepsilon)$, this error bound can be improved to $\mathcal{O}(\min\{\varepsilon, h\})$. The integrators thus do not resolve the $\mathcal{O}(\varepsilon)$ oscillations in η for large step sizes $h \ge \varepsilon$, but like in Jahnke & Lubich (2003)

they can be combined with a (symmetric and scaling-invariant) adaptive step size strategy such that the methods follow the non-adiabatic transitions through avoided crossings of eigenvalues with small steps and take large steps elsewhere.

We here consider applying an *integrating reversible step size controller* as in Sect. VIII.3.2 with the step size density function

$$\sigma(t) = \big(\|W(t)\|^2 + \alpha^2\big)^{-1/2}$$

for a parameter α that can be interpreted as the ratio of the accuracy parameter and the maximum admissible step size. Choosing the Frobenius norm $\|W\| = (\mathrm{trace}\, W^T W)^{1/2}$, we then obtain the following version of Algorithm VIII.3.4, where μ is the accuracy parameter and

$$G(t) = -\frac{\dot\sigma(t)}{\sigma(t)} = \big(\|W(t)\|^2 + \alpha^2\big)^{-1} \,\mathrm{trace}\,\big(\dot W(t)^T W(t)\big).$$

Set $z_0 = 1/\sigma(t_0)$ and, for $n \geq 0$,

$$\begin{aligned} z_{n+1/2} &= z_n + \frac{\mu}{2} G(t_n) \\ h_{n+1/2} &= \mu / z_{n+1/2} \\ t_{n+1} &= t_n + h_{n+1/2} \\ \eta_n &\mapsto \eta_{n+1} \quad \text{by (1.12) or (1.13) with step size } h_{n+1/2} \\ z_{n+1} &= z_{n+1/2} + \frac{\mu}{2} G(t_{n+1}). \end{aligned} \tag{1.14}$$

We remark that the schemes (1.12) and (1.13) can be modified such that they use evaluations at t_n and t_{n+1} instead of $t_{n+1/2}$ (Exercise 6).

Applying the above algorithm with accuracy parameter $\mu = 0.01$ and $\alpha = 0.1$ to the problem of Fig. 1.1 with $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ yields the step size sequences shown in Fig. 1.2. In each case the error at the end-point $t = 1$ was between $0.5 \cdot 10^{-3}$ and $2 \cdot 10^{-3}$.

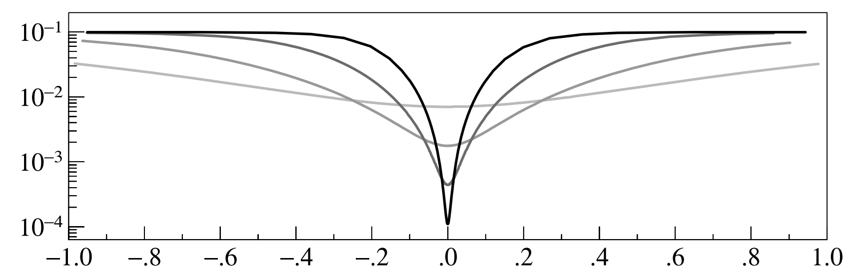


Fig. 1.2. Non-adiabatic transition: step sizes as function of t for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ (increasing darkness)

Second-Order Integrators. The $\mathcal{O}(\varepsilon)$ oscillations in η are resolved with step sizes up to $h = \mathcal{O}(\sqrt{\varepsilon})$ for methods that give $\mathcal{O}(h^2)$ accuracy uniformly in ε. Such

methods require a quadratic phase approximation, and one needs further terms obtained from reinserting $\eta(s)$ under the integral in (1.10) once again by the same formula, thus yielding terms with iterated integrals (this procedure is known as the *Neumann* or *Peano* or *Dyson* expansion in different communities, cf. Iserles 2004), or by including the first commutator in the *Magnus* expansion (IV.7.5). Symmetric second-order methods of both types are constructed by Jahnke (2004a).

Care must be taken in computing the arising oscillatory integrals. Iserles (2004) proposes and analyses Filon quadrature (after Filon, 1928), which is applicable when the moments, i.e., the integrals over products of oscillatory exponentials and polynomials, are known analytically. This is not the case, however, for all of the integrals appearing in the second-order methods. The alternative chosen by Jahnke (2004a) is to use an expansion technique based on partial integration. The idea can be illustrated on an integral such as

$$\int_0^1 \exp\Big(\frac{i\alpha\theta h}{\varepsilon}\Big) \cdot \exp\Big(\frac{i\beta\theta^2 h^2}{\varepsilon}\Big)\, d\theta$$

with $\alpha \neq 0$. Partial integration that integrates the first factor and differentiates the second factor yields a boundary term and again an integral of the same type, but now with an additional factor $\mathcal{O}\big(\frac{\varepsilon}{h}\cdot\frac{h^2}{\varepsilon}\big) = \mathcal{O}(h)$. Using this technique repeatedly in the oscillatory integrals appearing in the second-order methods permits to approximate all of them up to $\mathcal{O}(h^3)$ as needed. We refer to Jahnke (2004a) for the precise formulation and error analysis of these second-order methods, which are complicated to formulate, but do not require substantially more computational work than the first-order methods described above, and just the same number of matrix evaluations.

Higher-Order Integrators. Integrators of general order $p \geq 1$ are obtained with a phase approximation by polynomials of degree p and by including all terms of the Neumann or Magnus expansion for (1.5) with up to p-fold integrals.

XIV.2 Mechanical Systems with Time-Dependent Frequencies

We study oscillatory mechanical systems with explicitly time-dependent frequencies, where the time-dependent Hamiltonian is

$$H(p,q,t) = \frac{1}{2}\, p^T M(t)^{-1} p + \frac{1}{2\varepsilon^2}\, q^T A(t) q + U(q,t) \tag{2.1}$$

with a positive definite mass matrix $M(t)$ and a positive semi-definite stiffness matrix $A(t)$ of constant rank whose derivatives are bounded independently of ε. Such a Hamiltonian describes oscillations in a mechanical system that at the same time exerts a driven motion on a slower time scale. We consider motions of bounded energy:

$$H(p(t), q(t), t) \leq \text{Const.} \tag{2.2}$$

We transform (2.1) to a more amenable form by a series of linear time-dependent canonical coordinate transforms. The transformations turn the equations of motion into a form that approximately separates the time scales. This makes the problem more accessible to numerical discretization with large time steps and to the error analysis of multiple time-stepping methods applied directly to (2.1) in the originally given coordinates.

XIV.2.1 Canonical Transformation to Adiabatic Variables

By a series of canonical time-dependent linear transformations, which can all be done numerically with standard linear algebra routines, we now take the Hamiltonian system (2.1) to a form from which adiabatic invariants can be read off and which will serve as the base camp for both the construction and error analysis of numerical methods.

We introduce the energy E as the conjugate variable to time t and extend the Hamiltonian to

$$\widehat{H}(p, E, q, t) = H(p, q, t) + E. \tag{2.3}$$

The canonical equations of motion are then (the gradient ∇ refers only to q)

$$\begin{aligned} \dot{p} &= -\frac{1}{\varepsilon^2} A(t) q - \nabla U(q, t) \\ \dot{q} &= M(t)^{-1} p \end{aligned}$$

along with $\dot{E} = -\partial H / \partial t$ and $\dot{t} = 1$.

Transforming the Mass Matrix into the Identity Matrix. We change variables such that the mass matrix $M(t)$ in the kinetic energy part is replaced by the identity. With a smooth factorization

$$M(t)^{-1} = C(t) C(t)^T, \tag{2.4}$$

e.g., from a Cholesky decomposition of $M(t)$, we transform to variables $(\widetilde{q}, \widetilde{t})$ by

$$q = C(t)\widetilde{q}, \quad t = \widetilde{t}.$$

Then, the conjugate momenta are given by (see Example VI.5.2)

$$\begin{pmatrix} \widetilde{p} \\ \widetilde{E} \end{pmatrix} = \begin{pmatrix} C & \dot{C}\widetilde{q} \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} p \\ E \end{pmatrix} = \begin{pmatrix} C^T p \\ \widetilde{q}^T \dot{C}^T p + E \end{pmatrix}.$$

With the transformed matrix $\widetilde{A} = C^T A C$, the Hamiltonian $\widetilde{H}(\widetilde{p}, \widetilde{E}, \widetilde{q}, \widetilde{t}) = \widehat{H}(p, E, q, t)$ in the new variables then takes the form (we omit all tildes)

$$H(p, E, q, t) = \frac{1}{2} p^T p + \frac{1}{2\varepsilon^2} q^T A(t) q - q^T \dot{C}(t)^T C(t)^{-T} p + U(C(t)q, t) + E. \tag{2.5}$$

Diagonalizing the Stiffness Matrix. We diagonalize the matrix $A(t)$ in (2.5),

$$A(t) = Q(t) \begin{pmatrix} 0 & 0 \\ 0 & \Omega(t)^2 \end{pmatrix} Q(t)^T \tag{2.6}$$

with the diagonal matrix $\Omega(t) = \operatorname{diag}(\omega_j(t))$ of frequencies and an orthogonal matrix $Q(t)$, which depends smoothly on t if the frequencies remain separated. The matrix $Q(t)$ can be obtained as the product

$$Q(t) = Q_0(t) \begin{pmatrix} I & 0 \\ 0 & Q_*(t) \end{pmatrix}, \tag{2.7}$$

where the transformation with $Q_0(t)$ takes $A(t)$ to the block-diagonal form

$$A(t) = Q_0(t) \begin{pmatrix} 0 & 0 \\ 0 & A_*(t) \end{pmatrix} Q_0(t)^T$$

and $Q_*(t)$ diagonalizes $A_*(t)$. The effect of an avoided crossing of frequencies is localized to $Q_*(t)$, which then can have large derivatives, whereas those of $Q_0(t)$ remain moderately bounded. The transformation

$$q = Q(t)\widehat{q}, \quad t = \widehat{t}$$

with the conjugate momenta

$$\widehat{p} = Q(t)^T p, \quad \widehat{E} = \widehat{q}^T \dot{Q}(t)^T p + E$$

yields the Hamiltonian in the new variables $(\widehat{p}, \widehat{E}, \widehat{q}, \widehat{t})$ as (we omit all hats)

$$H = \frac{1}{2}\, p^T p + \frac{1}{2\varepsilon^2}\, q^T \begin{pmatrix} 0 & 0 \\ 0 & \Omega(t)^2 \end{pmatrix} q + q^T K(t) p + U(C(t)Q(t)q, t) + E \tag{2.8}$$

with

$$K = \begin{pmatrix} K_{00} & K_{01} \\ K_{10} & K_{11} \end{pmatrix} = Q^T \dot{Q} - Q^T \dot{C}^T C^{-T} Q.$$

We decompose also

$$p = \begin{pmatrix} p_0 \\ p_1 \end{pmatrix}, \quad q = \begin{pmatrix} q_0 \\ q_1 \end{pmatrix}$$

according to the blocks in (2.6) and refer to q_0 and q_1 (p_0 and p_1) as the *slow* and *fast* positions (slow and fast momenta), respectively. With the energy bound (2.2) we have

$$p_1 = \mathcal{O}(1), \quad q_1 = \mathcal{O}(\varepsilon). \tag{2.9}$$

Rescaling Positions and Momenta. We transform

$$q_0 = \check{q}_0, \quad q_1 = \varepsilon^{1/2}\Omega^{-1/2}\check{q}_1, \quad t = \check{t}$$

with the conjugate momenta

$$\check{p}_0 = p_0, \quad \check{p}_1 = \varepsilon^{1/2}\Omega^{-1/2}p_1, \quad \check{E} = -\frac{1}{2}\check{q}_1^T\varepsilon^{1/2}\Omega^{-3/2}\dot{\Omega}p_1 + E.$$

In the new variables, the Hamiltonian becomes (we omit the hačeks on all variables)

$$\begin{aligned} H &= \frac{1}{2}\,p_0^T p_0 + \frac{1}{2\varepsilon}\,p_1^T\Omega(t)p_1 + \frac{1}{2\varepsilon}\,q_1^T\Omega(t)q_1 \\ &\quad + q^T\check{K}(t)p + U(T(t)q,t) + E \end{aligned} \tag{2.10}$$

with

$$\begin{aligned} \check{K} &= \begin{pmatrix} K_{00} & \varepsilon^{-1/2}K_{01}\Omega^{1/2} \\ \varepsilon^{1/2}\Omega^{-1/2}K_{10} & \Omega^{-1/2}K_{11}\Omega^{1/2} + \frac{1}{2}\Omega^{-1}\dot{\Omega} \end{pmatrix} \\ T &= \begin{pmatrix} T_0 \mid \varepsilon^{1/2}T_1 \end{pmatrix} = \begin{pmatrix} T_{00} & \varepsilon^{1/2}T_{01} \\ T_{10} & \varepsilon^{1/2}T_{11} \end{pmatrix} = CQ\begin{pmatrix} I & 0 \\ 0 & \varepsilon^{1/2}\Omega^{-1/2} \end{pmatrix}. \end{aligned}$$

Eliminating the Singular Block. We next remove the $\mathcal{O}(\varepsilon^{-1/2})$ off-diagonal block in $\check{K}$ by the canonical transformation

$$-p_1 = -\overline{p}_1 + \varepsilon^{1/2}\Omega^{-1/2}K_{01}^T q_0, \quad q_0 = \overline{q}_0, \quad t = \overline{t}$$

with the conjugate variables

$$\overline{q}_1 = q_1, \quad \overline{p}_0 = p_0 + \varepsilon^{1/2}K_{01}\Omega^{-1/2}q_1, \quad \overline{E} = E + \varepsilon^{1/2}q_0^T\frac{d}{dt}\big(K_{01}\Omega^{-1/2}\big)q_1.$$

In these coordinates, the Hamiltonian takes the form (we omit all bars)

$$\begin{aligned} H &= \frac{1}{2}\,p_0^T p_0 + \frac{1}{2\varepsilon}\,p_1^T\Omega(t)p_1 + \frac{1}{2\varepsilon}\,q_1^T\Omega(t)q_1 \\ &\quad + q^T L(t)p + \frac{1}{2}\,q^T S(t)q + U(T(t)q,t) + E \end{aligned} \tag{2.11}$$

with the lower block-triangular matrix

$$\begin{aligned} L &= \begin{pmatrix} L_{00} & 0 \\ \varepsilon^{1/2}L_{10} & L_{11} \end{pmatrix} \\ &= \begin{pmatrix} K_{00} & 0 \\ \varepsilon^{1/2}\Omega^{-1/2}(K_{10}+K_{01}^T) & \Omega^{-1/2}K_{11}\Omega^{1/2} + \frac{1}{2}\Omega^{-1}\dot{\Omega} \end{pmatrix} \end{aligned}$$

and the symmetric matrix

$$S = \begin{pmatrix} S_{00} & \varepsilon^{1/2} S_{01} \\ \varepsilon^{1/2} S_{10} & \varepsilon S_{11} \end{pmatrix},$$

where

$$\begin{aligned} S_{00} &= -K_{01} K_{01}^T, \\ S_{01} &= S_{10}^T = -K_{00} K_{01} \Omega^{-1/2} \\ &\quad - K_{01} \Omega^{-1/2} (\Omega^{1/2} K_{11}^T \Omega^{-1/2} + \frac{1}{2} \Omega^{-1} \dot{\Omega}) - \frac{d}{dt} \big(K_{01} \Omega^{-1/2} \big), \\ S_{11} &= \Omega^{-1/2} (-K_{10} K_{01} - K_{01}^T K_{10}^T + K_{01}^T K_{01}) \Omega^{-1/2}. \end{aligned}$$

We note that with the energy bound (2.2) we now have

$$p_1 = \mathcal{O}(\varepsilon^{1/2}), \quad q_1 = \mathcal{O}(\varepsilon^{1/2}). \tag{2.12}$$

Equations of Motion. The differential equations now take the form

$$\begin{aligned} \dot{p}_0 &= f_0(p, q, t) \\ \dot{q}_0 &= p_0 + g_0(q, t) \\ \begin{pmatrix} \dot{p}_1 \\ \dot{q}_1 \end{pmatrix} &= \frac{1}{\varepsilon} \begin{pmatrix} 0 & -\Omega(t) \\ \Omega(t) & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} + \begin{pmatrix} f_1(p, q, t) \\ g_1(q, t) \end{pmatrix} \end{aligned} \tag{2.13}$$

with the functions bounded uniformly in ε,

$$\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = -L(t) p - S(t) q - T(t)^T \nabla U(T(t) q, t), \quad \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} = L(t)^T q.$$

The matrix in the system is diagonalized by a constant unitary matrix: with

$$\Gamma = \frac{1}{\sqrt{2}} \begin{pmatrix} I & I \\ -iI & iI \end{pmatrix} \tag{2.14}$$

we have

$$\begin{pmatrix} 0 & -\Omega(t) \\ \Omega(t) & 0 \end{pmatrix} = \Gamma \begin{pmatrix} i\Omega(t) & 0 \\ 0 & -i\Omega(t) \end{pmatrix} \Gamma^*. \tag{2.15}$$

Remark. Action-angle variables $p_{1,j} = \sqrt{a_j} \cos \theta_j$, $q_{1,j} = \sqrt{a_j} \sin \theta_j$ for the harmonic oscillators would now put the Hamiltonian into the form $H = \frac{1}{\varepsilon} \omega(t) \cdot a + G(a, \theta, p_0, q_0, t)$, which could be studied further using averaging techniques, that is, using coordinate transforms that reduce the dependence on the angles in the Hamiltonian; see Neishtadt (1984) for averaging out up to an exponentially small remainder in the case of a single high frequency. The first-order averaging transform might be done numerically (cf. the formulas in Sect. XII.2), but the higher-order transforms involve increasingly higher derivatives of the functions involved and therefore become impractical from the numerical viewpoint. For systems with several frequencies the averaging transforms require multi-dimensional integrals which are

expensive to compute. For our numerical purposes we therefore continue differently, adapting the adiabatic transformation of Sect. XIV.1.1.

The System in Adiabatic Variables. Let the diagonal phase matrix be given as

$$\Phi(t) = \int_{t_0}^{t} \Lambda(s)\, ds \qquad \text{with} \quad \Lambda(t) = \begin{pmatrix} \Omega(t) & 0 \\ 0 & -\Omega(t) \end{pmatrix}.$$

Our final transformation follows (1.4) and sets

$$\eta = \varepsilon^{-1/2} \exp\Bigl(-\frac{i}{\varepsilon}\Phi(t)\Bigr)\, \Gamma^* \begin{pmatrix} p_1 \\ q_1 \end{pmatrix}. \tag{2.16}$$

The factor $\varepsilon^{-1/2}$ is chosen for convenience so that (2.12) implies

$$\eta = \mathcal{O}(1). \tag{2.17}$$

We remark that up to now all transformations were invariant under rescaling $\varepsilon \to \sigma\varepsilon$ and $A(t) \to \sigma^2 A(t)$, but here we have chosen to give up this invariance in favour of (2.17). Note that η is of the form

$$\eta = \varepsilon^{-1/2}\Gamma^* \begin{pmatrix} \pi \\ \rho \end{pmatrix} = \frac{\varepsilon^{-1/2}}{\sqrt{2}} \begin{pmatrix} \pi + i\rho \\ \pi - i\rho \end{pmatrix} \tag{2.18}$$

with real vectors π, ρ satisfying

$$\pi + i\rho = \exp\Bigl(-\frac{i}{\varepsilon}\int_{t_0}^{t} \Omega(s)\, ds\Bigr)(p_1 + iq_1). \tag{2.19}$$

We denote the inverse transform as

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \varepsilon^{1/2} \begin{pmatrix} P_1(t) \\ Q_1(t) \end{pmatrix} \eta \qquad \text{with} \quad \begin{pmatrix} P_1(t) \\ Q_1(t) \end{pmatrix} = \Gamma \exp\Bigl(\frac{i}{\varepsilon}\Phi(t)\Bigr). \tag{2.20}$$

Together with $e = E + \frac{1}{2\varepsilon}\, p_1^T \Omega(t) p_1 + \frac{1}{2\varepsilon}\, q_1^T \Omega(t) q_1$ and unaltered p_0, q_0, t this yields a canonical transformation $(p_0, \pi, e, q_0, \rho, t) \mapsto (p_0, p_1, E, q_0, q_1, t)$. The Hamiltonian reads in these variables

$$H = \frac{1}{2}\, p_0^T p_0 + q^T L(t) p + \frac{1}{2}\, q^T S(t) q + U(T(t)q, t) + e,$$

where on the right-hand side the components p_1, q_1 are expressed in terms of η and π, ρ by (2.20) and (2.18). The equations of motion now become

$$\begin{aligned} \dot p_0 &= f_0(p, q, t) \\ \dot q_0 &= p_0 + g_0(q, t) \\ \dot\eta &= \varepsilon^{-1/2} \exp\Bigl(-\frac{i}{\varepsilon}\Phi(t)\Bigr)\, \Gamma^* \begin{pmatrix} f_1(p, q, t) \\ g_1(q, t) \end{pmatrix} \end{aligned}$$

with p_1, q_1 expressed in terms of η by (2.20). Written out, the differential equations for p_0, q_0 read

$$\begin{aligned}\dot{p}_0 &= -L_{00}p_0 - S_{00}q_0 - T_0^T \nabla U(T_0 q_0, t) - \varepsilon S_{01} Q_1 \eta \\ &\quad -T_0^T \Big(\nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t) - \nabla U(T_0 q_0, t) \Big) \\ \dot{q}_0 &= p_0 + L_{00}^T q_0 + \varepsilon L_{10}^T Q_1 \eta. \end{aligned} \tag{2.21}$$

The matrix multiplying η after substituting the expressions f_1 and g_1 in the differential equation for η becomes, apart from the oscillatory exponentials,

$$\begin{aligned} W &= \Gamma^* \begin{pmatrix} -L_{11} & -\varepsilon S_{11} \\ 0 & L_{11}^T \end{pmatrix} \Gamma \qquad (2.22) \\ &= -\frac{1}{2} \begin{pmatrix} L_{11} - L_{11}^T & L_{11} + L_{11}^T \\ L_{11} + L_{11}^T & L_{11} - L_{11}^T \end{pmatrix} - \frac{i\varepsilon}{2} \begin{pmatrix} -S_{11} & S_{11} \\ -S_{11} & S_{11} \end{pmatrix}, \end{aligned}$$

which has a diagonal of size $\mathcal{O}(\varepsilon)$. The equation for η then reads

$$\begin{aligned} \dot{\eta} &= \exp\Big(-\frac{i}{\varepsilon}\Phi(t)\Big)\, W(t)\, \exp\Big(\frac{i}{\varepsilon}\Phi(t)\Big)\, \eta \\ &\quad - P_1^* \Big(L_{10} p_0 + S_{10} q_0 + T_1^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t) \Big). \end{aligned} \tag{2.23}$$

The matrix multiplying η is bounded independently of ε, but highly oscillatory. Note that the coordinate transforms leading to (2.21), (2.23) are linear and can be carried out by standard numerical linear algebra routines.

Adiabatic Invariants. We suppose that the eigenfrequencies $\omega_j(t)$ remain separated and bounded away from 0: there are $\delta > 0$ and $c > 0$ such that for any pair $\omega_j(t)$ and $\omega_k(t)$ with $j \neq k$ ($j, k = 1, \ldots, m$), the lower bounds

$$|\omega_j(t) - \omega_k(t)| \geq \delta, \qquad \omega_j(t) \geq c \tag{2.24}$$

hold for all t under consideration. Under condition (2.24) the right-hand side $r(t)$ in the differential equation for η consists only of oscillatory terms, up to $\mathcal{O}(\varepsilon)$. (No smooth terms larger than $\mathcal{O}(\varepsilon)$ arise because the matrix W has a diagonal of size $\mathcal{O}(\varepsilon)$.) It then follows by partial integration that

$$\int_0^t r(s)\, ds = \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \leq \text{Const.}, \tag{2.25}$$

and as in (1.6) we then obtain

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \leq \text{Const.} \tag{2.26}$$

The functions defined by

$$I_j = |\eta_j|^2 \qquad (j = 1, \ldots, m) \tag{2.27}$$

are thus adiabatic invariants:

$$I_j(t) = I_j(0) + \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \le \text{Const.} \tag{2.28}$$

Starting from a Hamiltonian system (2.1), where the mass matrix equals the identity and the stiffness matrix is already diagonal, we find that I_j is the action (energy divided by frequency)

$$I_j(t) = \frac{1}{\omega_j(t)} \Big(\frac{1}{2}\, p_j(t)^2 + \frac{\omega_j(t)^2}{2\varepsilon^2}\, q_j(t)^2 \Big),$$

which for a constant frequency ω_j becomes a constant multiple of the oscillatory energy considered in Sect. XIII.9.

The Slow Limit System. As $\varepsilon \to 0$, the evolution of the slow variables p_0, q_0 is governed by the equations

$$\begin{aligned} \dot p_0 &= -L_{00}(t)p_0 - S_{00}(t)q_0 - T_0(t)^T \nabla U(T_0(t)q_0, t) \\ \dot q_0 &= p_0 + L_{00}(t)^T q_0 \end{aligned} \tag{2.29}$$

which is the system with the time-dependent Hamiltonian

$$H_0(p_0, q_0, t) = \frac{1}{2}\, p_0^T p_0 + q_0^T L_{00}(t) p_0 + \frac{1}{2}\, q_0^T S_{00}(t) q_0 + U(T_0(t)q_0, t).$$

We conclude this subsection with a simple illustration of the above procedure.

Example 2.1 (Harmonic oscillator with slowly varying frequency). For the scalar second-order differential equation

$$\ddot q + \frac{\omega(t)^2}{\varepsilon^2}\, q = 0,$$

where $\omega(t)$ is bounded away from 0 and has a derivative bounded independently of ε, the above transformations simplify considerably. The Hamiltonian in the original variables is already of the form

$$H = \frac{1}{2} p^2 + \frac{1}{2} \frac{\omega(t)^2}{\varepsilon^2}\, q^2,$$

and hence the first two transformations are not needed at all, and there are no slow variables p_0, q_0. The rescaling transformation yields the Hamiltonian (2.10) in the form

$$H = \frac{\omega(t)}{2\varepsilon} \check p^2 + \frac{\omega(t)}{2\varepsilon} \check q^2 + \frac{1}{2} \frac{\dot\omega(t)}{\omega(t)} \check p \check q.$$

With the adiabatic transformation (2.19) we thus represent the solution as

$$\sqrt{\frac{\varepsilon}{\omega(t)}}\, \dot q(t) + i \sqrt{\frac{\omega(t)}{\varepsilon}}\, q(t) = \exp\Big(\frac{i}{\varepsilon} \int_{t_0}^t \omega(s)\, ds \Big) \zeta(t),$$

where $\zeta = \pi + i\rho$ solves the differential equation

$$\dot{\zeta}(t) = -\frac{1}{2}\frac{\dot{\omega}(t)}{\omega(t)}\, \exp\Big(-\frac{2i}{\varepsilon}\int_{t_0}^{t}\omega(s)\,ds\Big)\zeta(t)$$

and satisfies $\zeta(t) = \zeta(t_0)\big(1+\mathcal{O}(\varepsilon)\big)$ for $t = \mathcal{O}(1)$. (In the above notation, we have $\eta = \frac{1}{\sqrt{2}}\varepsilon^{-1/2}(\zeta, \overline{\zeta})^T$.) The action

$$I(t) = \frac{1}{\omega(t)}\,\Big(\frac{1}{2}\,\dot{q}(t)^2 + \frac{\omega(t)^2}{2\varepsilon^2}\,q(t)^2\Big)$$

is an adiabatic invariant.

XIV.2.2 Adiabatic Integrators

A simple long-time-step integrator for the oscillatory mechanical system with time-dependent Hamiltonian (2.1) now reads as follows:

– Solve the slow limit system (2.29) for p_0, q_0, e.g., by the Störmer-Verlet method.
– Keep the adiabatic variable η constant at its initial value.

Under the condition of bounded energy (2.1) and the frequency separation condition (2.24), the error in η is then $\mathcal{O}(\varepsilon)$ over intervals $t \le$ Const. by (2.26). The difference between the solutions of (2.21) and the limit equation (2.29) is bounded by $\mathcal{O}(\varepsilon^2)$ for $t \le$ Const., as can be shown by forming the difference of the equations, integrating, estimating the integral of the extra terms by $\mathcal{O}(\varepsilon^2)$ using (2.26) and partial integration, and applying the Gronwall inequality. In the original variables p, q of (2.1) this yields an error $\mathcal{O}(\varepsilon^2)$ in the positions and $\mathcal{O}(\varepsilon)$ in the momenta.

More refined integrators are needed for two independent reasons:

1. to keep control of η on subintervals where the frequencies are not well separated and where η may thus deviate from its near-constant value;
2. to obtain higher order of approximation on intervals with separated frequencies.

We simplify the following presentation by assuming that the potential U is quadratic:

$$U(q,t) = \frac{1}{2}q^T G(t) q,$$

with a symmetric matrix $G(t)$ depending smoothly on t. We leave the required modifications for general U to the interested reader. Alternatively, the method with $U = 0$ can be used in the splitting approach of Sect. XIV.2.3 below.

An adiabatic integrator as described in Sect. XIV.1.2 can be extended to (2.23) and combined with a symmetric splitting between the weakly coupled systems (2.21) and (2.23): we begin with a symplectic Euler half-step for p_0, q_0 (denoting the time levels by superscripts),

$$\begin{aligned} p_0^{1/2} &= p_0^0 - \frac{h}{2}\Big(L_{00} p_0^{1/2} + (S_{00} + T_0^T G T_0) q_0^0 \\ &\qquad + \varepsilon \big(S_{01} + T_0^T G T_1 \big) \mathcal{Q}_1^- \eta^0 \Big) \qquad (2.30) \\ q_0^{1/2} &= q_0^0 + \frac{h}{2}\Big(p_0^{1/2} + L_{00}^T q_0^0 + \varepsilon L_{10}^T \mathcal{Q}_1^- \eta^0 \Big). \end{aligned}$$

Here the matrix functions L_{00}, L_{10}, S_{00}, S_{01}, T_0, T_1 are evaluated at $t_{1/2} = t_0 + h/2$, and $\mathcal{Q}_1^-$ is the average of the oscillatory function Q_1 of (2.20) over the half-step,

$$\mathcal{Q}_1^- \approx \frac{2}{h} \int_{t_0}^{t_{1/2}} Q_1(t)\, dt,$$

obtained with a linear approximation of the phase $\Phi(t)$ and analytic computation of the integral. We then make a full step for η with Eq. (2.23) like in (1.12),

$$\begin{aligned} \eta^1 &= \eta^0 + h\Big(E(\Phi) \bullet \mathcal{I} \bullet W \Big) \frac{1}{2} (\eta^1 + \eta^0) \\ &\qquad - h \mathcal{P}_1^* \Big(L_{10} p_0^{1/2} + (S_{10} + T_1^T G T_0) q_0^{1/2} \Big), \qquad (2.31) \end{aligned}$$

where again all matrix functions are evaluated at $t_{1/2}$, and $\mathcal{P}_1$ is the linear-phase approximation to the average

$$\mathcal{P}_1 \approx \frac{1}{h} \int_{t_0}^{t_1} P_1(t)\, dt.$$

The matrix W is as in (2.22), but with S_{11} replaced by $S_{11} + T_1^T G T_1$. The step is completed by a half-step for p_0, q_0 with the adjoint symplectic Euler method:

$$\begin{aligned} p_0^1 &= p_0^{1/2} - \frac{h}{2}\Big(L_{00} p_0^{1/2} + (S_{00} + T_0^T G T_0) q_0^1 \\ &\qquad + \varepsilon \big(S_{01} + T_0^T G T_1 \big) \mathcal{Q}_1^+ \eta^1 \Big) \qquad (2.32) \\ q_0^1 &= q_0^{1/2} + \frac{h}{2}\Big(p_0^{1/2} + L_{00}^T q_0^1 + \varepsilon L_{10}^T \mathcal{Q}_1^+ \eta^1 \Big), \end{aligned}$$

where the matrix functions are still evaluated at $t_{1/2}$, and $\mathcal{Q}_1^+$ approximates the average of Q_1 over the second half-step.

We now give local error bounds for this integrator, under conditions that include the case of an avoided crossing of frequencies.

Theorem 2.2. *Suppose that the functions in (2.1) are smooth and the frequencies satisfy (2.24) with minimal distance $\delta > 0$ for $t_0 \le t \le t_0 + h$, and the orthogonal matrix $Q_*(t)$ of (2.7), which diagonalizes the nonsingular part of the stiffness matrix, has derivatives bounded by $\dot{Q}_*(t) = \mathcal{O}(\delta^{-1})$, $\ddot{Q}_*(t) = \mathcal{O}(\delta^{-2})$. Assume further the energy bound (2.2) for the initial values. Then, the local error of method (2.30)–(2.32) is bounded by*

$$\begin{aligned} p_0^1 - p_0(t_0+h) &= \mathcal{O}(h^3/\delta^2) + \mathcal{O}(\varepsilon h^2/\delta^2) \\ q_0^1 - q_0(t_0+h) &= \mathcal{O}(h^3/\delta) + \mathcal{O}(\varepsilon h^2/\delta^2) \\ \eta^1 - \eta(t_0+h) &= \mathcal{O}(h^2/\delta^2). \end{aligned}$$

The constants symbolized by $\mathcal{O}$ do not depend on ε, h, and δ.

Proof. (a) Under the given conditions we have

$$K_{00} = \mathcal{O}(1),\ \ K_{01} = \mathcal{O}(1),\ \ K_{10} = \mathcal{O}(1),\ \ K_{11} = \mathcal{O}(\delta^{-1}),\ \text{and}$$

$$\dot K_{00} = \mathcal{O}(1),\ \ \dot K_{01} = \mathcal{O}(\delta^{-1}),\ \ \dot K_{10} = \mathcal{O}(\delta^{-1}),\ \ \dot K_{11} = \mathcal{O}(\delta^{-2}),$$

This yields the bounds

$$L_{00}, L_{10}, S_{00}, S_{11} = \mathcal{O}(1)$$

and similarly for their derivatives, and

$$L_{11}, S_{01}, S_{10} = \mathcal{O}(\delta^{-1}), \quad \dot L_{11}, \dot S_{01}, \dot S_{10} = \mathcal{O}(\delta^{-2}),$$

and hence also

$$W = \mathcal{O}(\delta^{-1}), \quad \dot W = \mathcal{O}(\delta^{-2}).$$

So we have from the energy bound and the differential equation (2.23) for η,

$$\eta = \mathcal{O}(1), \quad \dot\eta = \mathcal{O}(\delta^{-1}).$$

From the differential equations (2.21) for p_0, q_0 we conclude

$$\ddot p_0 = \mathcal{O}(\delta^{-1}) + \mathcal{O}(\varepsilon\delta^{-2}), \qquad \ddot q_0 = \mathcal{O}(\varepsilon\delta^{-1}).$$

(b) To study the local error in η, we integrate (2.23) from t_0 to t_0+h and compare with the corresponding term in (2.31):

$$\begin{aligned} &\int_{t_0}^{t_0+h} P_1^*(t)\Big(L_{10}p_0 + (S_{10} + T_1^T G T_0)q_0\Big)(t)\,dt \\ &- h\mathcal{P}_1^*\Big(L_{10}(t_{1/2})p_0^{1/2} + (S_{10} + T_1^T G T_0)(t_{1/2})q_0^{1/2}\Big) \\ &= \ \mathcal{O}(h^2/\delta^2), \end{aligned}$$

where we have used the above bounds and the error estimate for the linear phase approximation in the average of $P_1(t)$, cf. Sect. XIV.1.2,

$$\mathcal{P}_1 - \frac{1}{h}\int_{t_0}^{t_1} P_1(t)\,dt = \mathcal{O}(h/\delta).$$

Combining this estimate with the error bound of the adiabatic midpoint rule for the homogeneous equation as given in Theorem 1.2 yields the stated error bound for η_1.

(c) The error bound for the components p_0, q_0 comes about by combining error bounds for the Störmer–Verlet method (which require the bounds for $\ddot{p}_0, \ddot{q}_0$) and the estimates

$$\int_{t_0}^{t_0+h/2} \varepsilon (S_{01} + T_0^T G T_1) Q_1 \eta(t)\, dt - \frac{h}{2}\, \varepsilon \big(S_{01} + T_0^T G T_1\big)(t_{1/2}) \mathcal{Q}_1^- \eta^0$$
$$= \mathcal{O}(\varepsilon h^2/\delta^2)$$

and

$$\int_{t_0}^{t_0+h/2} \varepsilon L_{10}^T Q_1 \eta(t)\, dt - \frac{h}{2}\, \varepsilon L_{10}(t_{1/2}) \mathcal{Q}_1^- \eta^0 = \mathcal{O}(\varepsilon h^2/\delta),$$

and the same estimates for the second half-step. See also Exercise 7 for a similar situation. □

In the case of well-separated eigenvalues, the global error on bounded time intervals is thus bounded by $\mathcal{O}(h^2) + \mathcal{O}(h\varepsilon)$ in p_0, q_0 for $t \leq$ Const. and by $\mathcal{O}(h)$ in η. In the original variables p, q of (2.1), this then yields an error

$$q_n - q(t_n) = \mathcal{O}(h^2) + \mathcal{O}(h\varepsilon), \qquad p_n - p(t_n) = \mathcal{O}(h) \qquad \text{for} \quad t_n \leq \text{Const.}$$

With an adaptive step size strategy as in Sect. XIV.1.2, it is again possible to follow η through non-adiabatic transitions near avoided crossings of eigenvalues.

A higher-order scheme with a global error of $\mathcal{O}(h^2)$ in η – in the situation of separated eigenvalues – is obtained by replacing the upper line in (2.31) by a second-order adiabatic integrator as discussed in Sect. XIV.1.2, leaving the last term in (2.31) unaltered. In the original variables p, q of (2.1), the error is then $\mathcal{O}(h^2)$ both in positions and (fast and slow) momenta. The error is even $\mathcal{O}(\varepsilon h^2)$ in the fast positions q_1 of (2.8), which oscillate with an amplitude $\mathcal{O}(\varepsilon)$. We refer to Lorenz, Jahnke & Lubich (2005) for the particular case of second-order differential equations $\ddot{q} + \varepsilon^{-2} A(t) q = 0$ with a positive definite matrix $A(t)$.

XIV.2.3 Error Analysis of the Impulse Method

The transformation to adiabatic variables of Sect. XIV.2.1 also gives new insight into the error behaviour of multiple time stepping methods such as the impulse or mollified impulse method discussed in Sections VIII.4 and XIII.1, which do not use coordinate transforms in the method formulation. These methods are of interest when the eigendecompositions needed in adiabatic integrators are computationally more expensive than doing many small steps with the fast subsystem, and when evaluations of the potential force are so costly that the computational work for the fast subsystem becomes irrelevant. We consider the splitting

$$H = H^{\text{fast}} + H^{\text{slow}}$$

of the Hamiltonian (2.3) with

$$\begin{aligned}
H^{\text{fast}}(p,E,q,t) &= \frac{1}{2}\,p^T M(t)^{-1}p + \frac{1}{2\varepsilon^2}\,q^T A(t)q + E \\
H^{\text{slow}}(p,E,q,t) &= U(q,t).
\end{aligned}$$

The impulse method is given as the composition of the exact flows of the subsystems (see Sections VIII.4 and XIII.1.3):

$$\Phi_h = \varphi^{\text{slow}}_{h/2} \circ \varphi^{\text{fast}}_h \circ \varphi^{\text{slow}}_{h/2}\,,$$

where we are interested in taking long time steps $h \ge c\,\varepsilon$ (with a positive constant c). The equations of motion of the slow subsystem,

$$\dot p = -\nabla U(q,t), \quad \dot q = 0, \quad \dot t = 0,$$

are solved trivially by

$$\widehat p = p - \frac{h}{2}\nabla U(q,t), \quad \widehat q = q, \quad \widehat t = t.$$

In contrast, the fast subsystem needs to be integrated approximately, e.g., by many small substeps with the Störmer–Verlet method in the original variables (p,q) or by one step of the method (2.30)–(2.32) with $G=0$ in adiabatic variables (p_0,q_0,η). In the following we ignore the error resulting from this additional approximation and study the splitting method with exact flows.

The error behaviour of this method can be understood with the help of the transformation to adiabatic variables of Sect. XIV.2.1. The impulse method in the adiabatic variables p_0,q_0,η is obtained by splitting the differential equations (2.21) and (2.23). The fast subsystem is obtained by simply putting $U=0$ in these equations, and the slow subsystem reads

$$\begin{aligned}
\dot p_0 &= -T_0^T\nabla U(T_0q_0 + \varepsilon T_1Q_1\eta, t), \qquad \dot q_0 = 0 \\
\dot \eta &= -P_1^*T_1^T\nabla U(T_0q_0 + \varepsilon T_1Q_1\eta, t)
\end{aligned}$$

along with $\dot t = 0$, so that the argument in all the matrices is frozen at the initial time. Here $P_1(t)$ and $Q_1(t)$ are again the highly oscillatory matrix functions of (2.20). Since $Q_1P_1^*=0$ we have $Q_1\eta=$ Const., and therefore, in these variables the flow $\varphi^{\text{slow}}_{h/2}$ is the mapping given by

$$\begin{aligned}
\widehat p_0 &= p_0 - \frac{h}{2}\,T_0^T\nabla U(T_0q_0 + \varepsilon T_1Q_1\eta, t_0), \qquad \widehat q_0 = q_0 \\
\widehat \eta &= \eta - \frac{h}{2}\,P_1^*T_1^T\nabla U(T_0q_0 + \varepsilon T_1Q_1\eta, t_0),
\end{aligned} \tag{2.33}$$

where the matrices T_0, T_1, P_1, Q_1 are evaluated at t_0. In the impulse method, the above values are the starting values for a step with φ^{fast}_h, which is followed by another application of $\varphi^{\text{slow}}_{h/2}$.

A disturbing feature in (2.33) is the appearance of the particular value $P_1(t_0)$ of the highly oscillatory function instead of the average $\mathcal{P}_1$ as in (2.31).

We now consider the error propagation for η in the case of well-separated frequencies. Recall that the exact solution then satisfies $\eta(t) = \eta(0) + \mathcal{O}(\varepsilon)$ for $t \le \mathrm{Const}$. For ease of presentation we consider a constant step size h.

Lemma 2.3. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the result after n steps satisfies, for $nh \le T \le$ Const.,*

$$\eta_n = \eta_0 + \sigma_n + \mathcal{O}(\varepsilon), \tag{2.34}$$

where

$$\|\sigma_n\| \le C\kappa \qquad \textit{with} \quad \kappa = \max_{0\le nh\le T} \max_k \Big\| h \sum_{j=0}^{n} \exp\Big(\frac{i}{\varepsilon}\,\phi_k(t_j)\Big)\Big\|. \tag{2.35}$$

Proof. We have $\eta_n = \eta_h(t_n)$, where $\eta_h(t)$ solves the differential equation with impulses,

$$\dot\eta_h = \exp\Big(-\frac{i}{\varepsilon}\Phi\Big)\, W \exp\Big(\frac{i}{\varepsilon}\Phi\Big)\,\eta_h + r + \sum_j \varDelta\eta_j\,\delta_j .$$

Here $W(t)$ is the matrix (2.22) appearing in (2.23), and

$$r(t) = -P_1^*(t)\big(L_{10}(t)p_{0,h}(t) + S_{01}(t)q_{0,h}(t)\big)$$

with $p_{0,h}(t)$, $q_{0,h}(t)$ denoting the piecewise constant functions that take the values of the numerical solution. Further we have

$$\varDelta\eta_j = -hP_1(t_j)^*\, T_1(t_j)^T\, \nabla U\big(T_0(t_j)q_{0,j} + \varepsilon T_1(t_j)Q_1(t_j)\eta_j, t_j\big),$$

the expression on the right-hand side of (2.33), and δ_j is a Dirac impulse located at t_j. It follows that, for $t = nh$,

$$\begin{aligned}\eta_n - \eta_0 &= \eta_h(t_n) - \eta_h(0)\\ &= \int_0^t \exp\Big(-\frac{i}{\varepsilon}\Phi(s)\Big)\, W(s)\, \exp\Big(\frac{i}{\varepsilon}\Phi(s)\Big)\,\eta_h(s)\,ds + \int_0^t r(s)\,ds + \sigma_n,\end{aligned}$$

where σ_n is the trapezoidal sum of the terms on the right-hand side of (2.33):

$$\sigma_n = -h{\sum_{j=0}^{n}}{}'\; P_1(t_j)^*\, T_1(t_j)^T\, \nabla U\big(T_0(t_j)q_{0,j} + \varepsilon T_1(t_j)Q_1(t_j)\eta_j, t_j\big). \tag{2.36}$$

The prime on the sum indicates that the first and last term are taken with the factor $\frac{1}{2}$. Using partial integration as in (1.6), we obtain

$$\int_0^t \exp\Big(-\frac{i}{\varepsilon}\Phi(s)\Big)\, W(s)\, \exp\Big(\frac{i}{\varepsilon}\Phi(s)\Big)\,\eta_h(s)\,ds = \mathcal{O}(\varepsilon),$$

and by partial integration as in (2.25),

$$\int_0^t r(s)\,ds = \mathcal{O}(\varepsilon).$$

This shows (2.34). A partial summation in (2.36), summing up the oscillatory terms $P_1(t_j)$ and differencing the smoother other terms, then yields (2.35). □

The size of κ of (2.35) depends on possible resonances between the step size and the frequencies, yielding κ between $\mathcal{O}(h)$ and $\mathcal{O}(1)$. For the error of the method we have the following.

Theorem 2.4. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the error of the impulse method after n steps with step size $h \geq c\,\varepsilon$ satisfies*

$$\begin{aligned} p_n - p(t_n) &= \mathcal{O}(\kappa) \\ q_n - q(t_n) &= \mathcal{O}(h^2) + \mathcal{O}(\varepsilon\kappa). \end{aligned}$$

The constants symbolized by $\mathcal{O}$ do not depend on ε, h and n with $nh \leq \mathrm{Const.}$

Proof. The error of size $\mathcal{O}(\kappa)$ in η immediately implies an error of size $\mathcal{O}(\kappa)$ in the actions $I_j = \frac{1}{2}|\eta_j|^2$, and an error of $\mathcal{O}(\kappa)$ in the fast momenta p_1 and of $\mathcal{O}(\varepsilon\kappa)$ in the fast positions q_1 of (2.9); recall the transformation (2.16) and the rescaling. In the slow components p_0, q_0 the method is a perturbed variant of the Störmer–Verlet method. The contribution of the perturbations $\varepsilon T_1 Q_1 \eta$ to the error is of size $\mathcal{O}(\varepsilon\kappa)$. This is seen by applying the simple lemma below with $y = (p_0, q_0)$ and

$$d_n = \begin{pmatrix} -hT_0(t_n)^T \nabla^2 U(T_0(t_n)q_{0,n}, t_n)\,\varepsilon T_1(t_n) Q_1(t_n)\eta_n \\ 0 \end{pmatrix} + \mathcal{O}(h^2\varepsilon)$$

and using partial summation of the d_n, summing up the oscillatory terms $Q_1(t_n)$ and differencing the other terms. □

Lemma 2.5. *Let $\Phi_h(y) = y + hF_h(y)$ be a one-step method where F_h has Lipschitz constant L. Consider the method and a perturbation,*

$$y_{n+1} = \Phi_h(y_n) \qquad \text{and} \qquad \widetilde{y}_{n+1} = \Phi_h(\widetilde{y}_n) + d_n,$$

with the same starting values $\widetilde{y}_0 = y_0$. Then, the difference is bounded by

$$\|\widetilde{y}_n - y_n\| \leq e^{nhL} \cdot \max_{0 \leq k \leq n-1} \Big\| \sum_{j=0}^{k} d_j \Big\|.$$

Proof. The result follows from

$$\widetilde{y}_n - y_n = h \sum_{j=0}^{n-1} \big(F_h(\widetilde{y}_n) - F_h(y_n)\big) + \sum_{j=0}^{n-1} d_j$$

with the discrete Gronwall inequality. □

XIV.2.4 Error Analysis of the Mollified Impulse Method

The problem with possible step-size resonances can be greatly alleviated by the mollified impulse method (see Sect. XIII.1.4) where the potential $U(q,t)$ is replaced by a modified potential $\overline{U}(q,t)$. A good choice is

$$\overline{U}(q,t) = U(\mathcal{A}(t)q,t) \qquad \text{with} \quad \mathcal{A}(t) = C(t)Q(t)\begin{pmatrix} I & 0 \\ 0 & \mathcal{S}(t) \end{pmatrix} Q(t)^T C(t)^{-1} \tag{2.37}$$

with C and Q of (2.4) and (2.6), and

$$\mathcal{S}(t) = \operatorname{sinc}\Big(\frac{h}{\varepsilon}\Omega(t)\Big) = \frac{1}{2h}\int_{-h}^{h} \exp\Big(\pm\frac{is}{\varepsilon}\,\Omega(t)\Big)\,ds.$$

A calculation shows that it replaces (2.33) by

$$\begin{aligned} \widehat{p}_0 &= p_0 - \frac{h}{2}\,T_0^T\nabla U(T_0q_0 + \varepsilon T_1\mathcal{Q}_1\eta, t_0), \qquad \widehat{q}_0 = q_0 \\ \widehat{\eta} &= \eta - \frac{h}{2}\,\mathcal{P}_1^* T_1^T\nabla U(T_0q_0 + \varepsilon T_1\mathcal{Q}_1\eta, t_0), \end{aligned} \tag{2.38}$$

with matrix functions evaluated at t_0, where $\mathcal{P}_1(t)$ and $\mathcal{Q}_1(t)$ are the linear-phase approximations to the average over the interval $[t-h, t+h]$ of P_1 and Q_1, respectively,

$$\begin{aligned} \mathcal{P}_1(t) = \mathcal{S}(t)P_1(t) &= \frac{1}{2h}\int_{t-h}^{t+h} P_1(s)\,ds + \mathcal{O}(h) \\ \mathcal{Q}_1(t) = \mathcal{S}(t)Q_1(t) &= \frac{1}{2h}\int_{t-h}^{t+h} Q_1(s)\,ds + \mathcal{O}(h). \end{aligned}$$

Therefore, (2.34) and (2.36) hold with the highly oscillatory $P_1(t_j)$ replaced by the averages $\mathcal{P}_1(t_j)$. Using a partial summation in (2.36) and noting that, for $t = nh \le$ Const.,

$$\Big\| h\sum_{j=1}^{n} \mathcal{P}_1(t_j)\Big\| = \Big\| \int_0^t P_1(s)\,ds\Big\| + \mathcal{O}(h) = \mathcal{O}(\varepsilon) + \mathcal{O}(h),$$

we obtain an estimate

$$\eta_n = \eta_0 + \mathcal{O}(h)$$

instead of the corresponding bound (2.34) with (2.35). This eliminates the bad effect of step size resonances (large κ) on the propagation in the fast variables over bounded time intervals $t \le$ Const. (though not on longer intervals, as we know from Chap. XIII). The more harmless effect of step size resonances on the slow variables, as visible in the term $\mathcal{O}(\varepsilon\kappa)$ in Theorem 2.4, is likewise reduced to $\mathcal{O}(\varepsilon h)$. We thus obtain the following improvement over the error bounds in Theorem 2.4.

Theorem 2.6. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the error of the above mollified impulse method after n steps with step size $h \geq c\,\varepsilon$ satisfies*

$$\begin{aligned} p_n - p(t_n) &= \mathcal{O}(h) \\ q_n - q(t_n) &= \mathcal{O}(h^2). \end{aligned}$$

The constants symbolized by $\mathcal{O}$ do not depend on ε, h and n with $nh \leq \mathrm{Const}$. □

A direct implementation of this method requires just the same matrix decompositions that are needed for the integrators in adiabatic variables. It is then reasonable to use one step of the adiabatic integrator of Sect. XIV.2.2 for solving the fast subsystem over a time step.

An alternative is to compute the average $\mathcal{A}(t)$ by small time steps from the linear differential equation with the Hamiltonian H^{fast}, as formulated in Sect. XIII.1.4. The method described here then corresponds to (XIII.1.18) with $c = 1$.

XIV.3 Mechanical Systems with Solution-Dependent Frequencies

We[2] consider the Hamiltonian

$$H(p,q) = \frac{1}{2}\,p^T M(q)^{-1} p + U(q) + \frac{1}{\varepsilon^2} V(q) \tag{3.1}$$

with a strong potential $\varepsilon^{-2} V(q)$ that penalizes some directions of motion. Analytical studies of this problem were done by Rubin & Ungar (1957), Takens (1980), and Bornemann (1998). In an alternative approach to these works, we here describe a transformation of the problem to adiabatic variables. This gives new insight into the solution behaviour and can be used as the starting point for the construction of long-time-step integrators. It also enables us to analyse the error of multiple time-stepping methods.

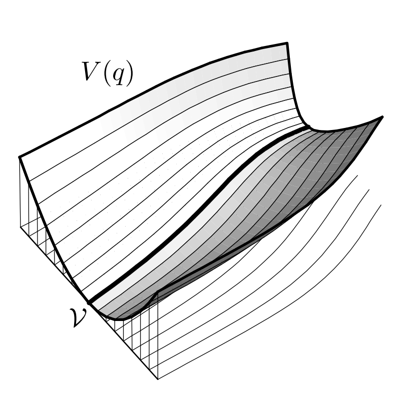
XIV.3.1 Constraining Potentials

We consider the Hamiltonian (3.1), where $M(q)$ is a symmetric positive definite mass matrix depending smoothly on the positions $q \in \mathbb{R}^n$, U is a smooth potential, and the constraining potential is assumed to satisfy the following:

[2] This section was written in cooperation with Katina Lorenz (Doctoral Thesis, Univ. Tübingen, in preparation).

The smooth function $V : D \subset \mathbb{R}^n \to \mathbb{R}$ attains its minimum value 0 on a d-dimensional manifold $\mathcal{V} \subset \mathbb{R}^n$,

$$\mathcal{V} = \{q \in D \mid V(q) = \min V = 0\}. \quad (3.2)$$



In a neighbourhood of $\mathcal{V}$, the potential V is strongly convex along directions non-tangential to $\mathcal{V}$, that is, there exists $\alpha > 0$ such that for $q \in \mathcal{V}$, the Hessian $\nabla^2 V(q)$ satisfies

$$v^T \nabla^2 V(q) v \geq \alpha \cdot v^T M(q) v \qquad (3.3)$$

for all vectors v in the $M(q)$-orthogonal complement of the tangent space $T_q\mathcal{V}$.

We let $m = n - d$ be the number of independent constraints that locally describe the manifold $\mathcal{V}$.

Example 3.1 (Chain of Stiff Springs). The position of $m + 1$ mass points in a plane, arranged in a chain connected by stiff springs with spring constants α_i^2/ε^2, is determined by the Cartesian coordinates of the first mass point and by m angles φ_i and the elongations d_i of the m springs. The constraining potential is

$$V = \frac{1}{2} \sum_{i=1}^{m} \alpha_i^2 \, d_i^2,$$

and the constraint manifold is described by $d_1 = \ldots = d_m = 0$ corresponding to non-elongated springs. The frequencies of the vibrations in such a chain depend on the angles.

In the above example we have, in the coordinates given by the angles and elongations, a potential V of the form

$$V(q) = \frac{1}{2} \, q_1^T A(q_0) q_1 \qquad (3.4)$$

for $q = (q_0, q_1) \in \mathbb{R}^d \times \mathbb{R}^m$, with a positive definite matrix $A(q_0)$. The manifold of constraints is here simply $\mathcal{V} = \mathbb{R}^d \times 0$. As the following lemma shows, this is already the general situation in suitable local coordinates.

Lemma 3.2. *Under conditions (3.2)–(3.3), there exists a smooth local change of coordinates $q = \chi(y)$ such that*

$$V(q) = \frac{1}{2} \, y_1^T A(y_0) y_1 \quad \textit{for} \quad q = \chi(y)$$

with $y = (y_0, y_1)$ near 0 in $\mathbb{R}^d \times \mathbb{R}^m$, where $A(y_0)$ is a symmetric positive definite $m \times m$ matrix.

Proof. In a first step, we choose local coordinates $q = \psi(x)$ with $x = (x_0, x_1)$ near 0 in $\mathbb{R}^d \times \mathbb{R}^m$, such that $q = \psi(x) \in \mathcal{V}$ if and only if $x_1 = 0$. In these coordinates, denoting $\widehat{V}(x) = V(q)$ for $q = \psi(x)$, we then have

$$\widehat{V}(x_0, 0) = 0, \quad \nabla \widehat{V}(x_0, 0) = 0$$

by (3.2), and

$$A(x_0) := \nabla_{x_1}^2 \widehat{V}(x_0, 0) \quad \text{is positive definite}$$

by (3.3). We now change coordinates by the near-identity transformation

$$y_0 = x_0, \quad y_1 = \mu(x) x_1$$

where the real factor $\mu(x)$ (near 1 for x_1 near 0) is to be chosen such that

$$\frac{1}{2} y_1^T A(y_0) y_1 = \widehat{V}(x_0, x_1).$$

Since the right-hand side equals

$$\widehat{V}(x_0, x_1) - \widehat{V}(x_0, 0) - x_1^T \nabla \widehat{V}(x_0, 0) = \frac{1}{2} x_1^T A(x_0) x_1 + r(x)$$

with $r(x) = \mathcal{O}(\|x_1\|^3)$, the choice

$$\mu(x) = \sqrt{1 + \frac{2r(x)}{x_1^T A(x_0) x_1}}$$

does the trick. □

We remark that Lemma 3.2 could be obtained as a corollary to the Morse lemma, for which we refer to Abraham & Marsden (1978) and Crouzeix & Rappaz (1989).

The change to the local coordinates $x = (x_0, x_1)$ such that $V(q) = 0$ if and only if $x_1 = 0$ for $q = \psi(x)$, is not numerically constructive from the mere knowledge of an expression for the potential V. However, in many situations the manifold $\mathcal{V}$ can be described by constraints $g(q) = 0$, and $x_1 = g$ can then be extended to a full set of coordinates. The above transformation from x to y can be done numerically. In the usual way, the transformation $q = \chi(y)$ of the position coordinates extends to a canonical transformation by setting $p_y = \chi'(y)^T p$ for the conjugate momenta; see Example VI.5.2.

Solutions of (3.1) are in general oscillatory with frequencies of size $\sim \varepsilon^{-1}$. There exist, however, special solutions having arbitrarily many time derivatives bounded independently of ε, which for arbitrary $N \geq 1$ stay $\mathcal{O}(\varepsilon^N)$ close to a manifold $\mathcal{V}^{\varepsilon,N}$ that has a distance $\mathcal{O}(\varepsilon)$ to $\mathcal{V}$. See Lubich (1993), where also implicit Runge-Kutta methods for the approximation of the smooth solutions are studied. In this section we are, however, interested in approximating general oscillatory solutions of bounded energy.

XIV.3.2 Transformation to Adiabatic Variables

We start from a Hamiltonian (3.1) in coordinates (p, q) where the constraining potential is already of the form (3.4) for $q = (q_0, q_1)$. We note that for a system of bounded energy, we then have $q_1 = \mathcal{O}(\varepsilon)$.

We now perform a series of canonical transformations that take the Hamiltonian into a form that is better suited for a direct numerical treatment and for the error analysis of multiple time-stepping methods. The transformations are similar to those for the time-dependent case treated in Sect. XIV.2.1, but here they appear in a permuted order.

Transforming the Stiffness Matrix into the Identity. We write the Cholesky decomposition of the stiffness matrix as

$$A(q_0) = C(q_0)^{-T} C(q_0)^{-1}$$

and change to variables

$$q_0 = \widetilde{q}_0, \qquad q_1 = C(\widetilde{q}_0)\widetilde{q}_1$$

along with the conjugate momenta

$$\widetilde{p}_0 = p_0 + \Big(\frac{\partial}{\partial \widetilde{q}_0} C(\widetilde{q}_0)\widetilde{q}_1\Big)^T p_1, \qquad \widetilde{p}_1 = C(\widetilde{q}_0)^T p_1.$$

With the transformed mass matrix $\widetilde{M}(\widetilde{q}) = B(\widetilde{q}) M(\widetilde{q}_0, C(\widetilde{q}_0)\widetilde{q}_1) B(\widetilde{q})^T$ (for the matrix $B(\widetilde{q})$ that transforms $\widetilde{p} = B(\widetilde{q})p$) and the potential $\widetilde{U}(\widetilde{q}) = U(\widetilde{q}_0, C(\widetilde{q}_0)\widetilde{q}_1)$, the Hamiltonian takes the simplified form (we omit all tildes)

$$H = \frac{1}{2}\, p^T M(q)^{-1} p + \frac{1}{2\varepsilon^2}\, q_1^T q_1 + U(q). \tag{3.5}$$

Eliminating Off-Diagonal Blocks in the Mass Matrix. We write the mass matrix $M(q)$ as

$$M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix}.$$

With $G(\overline{q}_0) = -M_{00}(\overline{q}_0, 0)^{-1} M_{01}(\overline{q}_0, 0)$, we transform

$$q_0 = \overline{q}_0 + G(\overline{q}_0)\overline{q}_1, \quad q_1 = \overline{q}_1,$$

with the conjugate momenta

$$\overline{p}_0 = p_0 + \Big(\frac{\partial}{\partial \overline{q}_0} G(\overline{q}_0)\overline{q}_1\Big)^T p_0, \quad \overline{p}_1 = p_1 + G(\overline{q}_0)^T p_0.$$

This canonical change of variables eliminates M_{01} and M_{10} in the transformed mass matrix $\overline{M}(q_0, 0)$ and keeps the Schur complement on the block diagonal: with the symmetric positive definite matrices

$$\overline{M}_0(\overline{q}_0) = M_{00}(\overline{q}_0, 0), \quad \overline{M}_1(\overline{q}_0) = \big(M_{11} - M_{10}M_{00}^{-1}M_{01}\big)(\overline{q}_0, 0),$$

the transformation puts the Hamiltonian into the form (we omit all bars)

$$\begin{aligned} H &= \frac{1}{2}\, p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2}\, p_1^T M_1(q_0)^{-1} p_1 + \frac{1}{2\varepsilon^2}\, q_1^T q_1 \\ &\quad + \frac{1}{2}\, p^T R(q) p + U(q_0 + G(q_0)q_1, q_1) \end{aligned} \tag{3.6}$$

where R is a smooth matrix-valued function satisfying

$$R(q_0, 0) = 0. \tag{3.7}$$

Diagonalizing the Mass Matrix of the Fast Variables. We diagonalize

$$M_1(q_0) = Q(q_0)\Omega(q_0)^{-2}Q(q_0)^T$$

with the diagonal matrix $\Omega(q_0) = \mathrm{diag}(\omega_j(q_0))$ of frequencies and an orthogonal matrix $Q(q_0)$, which depends smoothly on q_0 if the frequencies are separated. We transform

$$q_0 = \widehat{q}_0, \quad q_1 = Q(\widehat{q}_0)\widehat{q}_1$$

with the conjugate momenta

$$\widehat{p}_0 = p_0 + \Big(\frac{\partial}{\partial \widehat{q}_0}\, Q(\widehat{q}_0)\widehat{q}_1\Big)^T p_1, \quad \widehat{p}_1 = Q(\widehat{q}_0)^T p_1.$$

The matrix

$$Y(\widehat{q}) = \Big(\frac{\partial}{\partial \widehat{q}_0}\, Q(\widehat{q}_0)\widehat{q}_1\Big)^T Q(\widehat{q}_0)$$

is of size $\mathcal{O}(\widehat{q}_1)$ but it is this expression which may become large near avoided crossings of eigenvalues. We consider the associated matrix

$$X(\widehat{q}) = \begin{pmatrix} 0 & X_{01} \\ X_{10} & X_{11} \end{pmatrix} = \begin{pmatrix} 0 & -M_0^{-1}Y \\ -Y^T M_0^{-1} & Y^T M_0^{-1} Y \end{pmatrix}. \tag{3.8}$$

With a matrix $\widehat{R}(\widehat{q})$ satisfying (3.7), which is a sum of the appropriately transformed previous matrix R and the above matrix X, the Hamiltonian in the new variables $(\widehat{p}, \widehat{q})$ becomes (we omit all hats)

$$\begin{aligned} H &= \frac{1}{2}\, p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2}\, p_1^T \Omega(q_0)^2 p_1 + \frac{1}{2\varepsilon^2}\, q_1^T q_1 \\ &\quad + \frac{1}{2}\, p^T R(q) p + U(q_0 + GQ(q_0)q_1, Q(q_0)q_1). \end{aligned} \tag{3.9}$$

Rescaling Positions and Momenta. We change to rescaled fast variables

$$q_0 = \check{q}_0, \quad q_1 = \varepsilon^{1/2} \Omega(\check{q}_0)^{1/2} \check{q}_1$$

(note that $q_1 = \mathcal{O}(\varepsilon)$ implies $\check{q}_1 = \mathcal{O}(\varepsilon^{1/2})$) with the conjugate momenta

$$\check{p}_0 = p_0 + \varepsilon^{1/2} \Big(\frac{\partial}{\partial \check{q}_0} \, \Omega(\check{q}_0)^{1/2} \check{q}_1 \Big)^T p_1, \quad \check{p}_1 = \varepsilon^{1/2} \Omega(\check{q}_0)^{1/2} p_1 .$$

In the new variables, the Hamiltonian becomes (we omit the hačeks on all variables)

$$\begin{aligned} H &= \frac{1}{2} \, p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2\varepsilon} \, p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon} \, q_1^T \Omega(q_0) q_1 \\ &\quad + \frac{1}{2} \, p^T R(q) p + U(T(q_0) q), \end{aligned} \tag{3.10}$$

where

$$T = \Big(T_0 \mid \varepsilon^{1/2} T_1 \Big) = \begin{pmatrix} I & \varepsilon^{1/2} G Q \Omega^{1/2} \\ 0 & \varepsilon^{1/2} Q \Omega^{1/2} \end{pmatrix}$$

and $R(q)$ is a symmetric matrix of the form

$$R(q) = \begin{pmatrix} R_{00}(q_0, \varepsilon^{1/2} q_1) & \varepsilon^{-1/2} R_{01}(q_0, \varepsilon^{1/2} q_1) \\ \varepsilon^{-1/2} R_{10}(q_0, \varepsilon^{1/2} q_1) & \varepsilon^{-1} R_{11}(q_0, \varepsilon^{1/2} q_1) \end{pmatrix}$$

with smooth functions R_{ij} satisfying $R_{ij}(q_0, 0) = 0$. Therefore, the expression $\frac{1}{2} \, p^T R(q) p$ can be rewritten in the form

$$\begin{aligned} \frac{1}{2} \, p^T R(q) p &= \varepsilon^{1/2} c(p_0, q_0)^T q_1 + p_1^T L(p_0, q_0)^T q_1 \\ &\quad + \varepsilon^{-1/2} \tau(p_1, p_1, q_1; p_0, q_0) + \rho(p, q), \end{aligned} \tag{3.11}$$

with a vector c, a matrix L, a function τ that is trilinear in p_1, p_1, q_1, and a remainder of size $\rho(p, q) = \mathcal{O}(\varepsilon^2)$ for $p_1, q_1 = \mathcal{O}(\varepsilon^{1/2})$, whose partial derivatives with respect to p_1, q_1 are of size $\mathcal{O}(\varepsilon^{3/2})$, and with respect to p_0, q_0 of size $\mathcal{O}(\varepsilon^2)$.

Equations of Motion. The differential equations now take the form

$$\begin{aligned} \dot{p}_0 &= -\nabla_{q_0} \Big(\frac{1}{2} \, p_0^T M_0(q_0)^{-1} p_0 + U(q_0, 0) \Big) \\ &\quad - \nabla_{q_0} \Big(\frac{1}{2\varepsilon} \, p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon} \, q_1^T \Omega(q_0) q_1 \Big) + f_0(p, q) \\ \dot{q}_0 &= M_0(q_0)^{-1} p_0 + g_0(p, q) \\ \begin{pmatrix} \dot{p}_1 \\ \dot{q}_1 \end{pmatrix} &= \frac{1}{\varepsilon} \begin{pmatrix} 0 & -\Omega(q_0) \\ \Omega(q_0) & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} + \begin{pmatrix} f_1(p, q) \\ g_1(p, q) \end{pmatrix} \end{aligned} \tag{3.12}$$

with the functions

$$\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = -\nabla_q \Big(\frac{1}{2} p^T R(q) p + U(T(q_0)q) - U(q_0,0) \Big)$$
$$\begin{pmatrix} g_0 \\ g_1 \end{pmatrix} = R(q)p.$$

We note the magnitudes $f_0 = \mathcal{O}(\varepsilon)$, $g_0 = \mathcal{O}(\varepsilon)$ and $f_1 = \mathcal{O}(\varepsilon^{1/2})$, $g_1 = \mathcal{O}(\varepsilon^{1/2})$ in the case of separated eigenfrequencies, where the diagonalization is smooth with bounded derivatives. By (3.11) we have (omitting the arguments p_0, q_0 in c, L, T)

$$\begin{aligned} f_1 &= -\varepsilon^{1/2} c - L p_1 + \varepsilon^{-1/2} a(p_1, p_1; p_0, q_0) - \varepsilon^{1/2} T_1^T \nabla U(q_0, 0) + \mathcal{O}(\varepsilon^{3/2}) \\ g_1 &= L^T q_1 + \varepsilon^{-1/2} b(p_1, q_1; p_0, q_0) + \mathcal{O}(\varepsilon^{3/2}) \end{aligned} \tag{3.13}$$

where the functions a and b are bilinear in their first two arguments.

The System in Adiabatic Variables. We finally leave the canonical framework and transform to adiabatic variables as in (2.16). Along a solution $(p(t), q(t))$ of the system (3.12) we consider the diagonal phase matrix $\Phi(t)$ defined by

$$\dot{\Phi} = \Lambda(q_0) \quad \text{with} \quad \Lambda(q_0) = \begin{pmatrix} \Omega(q_0) & 0 \\ 0 & -\Omega(q_0) \end{pmatrix}.$$

With the constant unitary matrix Γ of (2.14), which diagonalizes the matrix in (3.12), we introduce the adiabatic variables

$$\eta = \varepsilon^{-1/2} \exp\Big(-\frac{i}{\varepsilon}\Phi\Big)\, \Gamma^* \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} \tag{3.14}$$

and denote the inverse transform as

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \varepsilon^{1/2} \begin{pmatrix} P_1 \\ Q_1 \end{pmatrix} \eta = \varepsilon^{1/2} \Gamma \exp\Big(\frac{i}{\varepsilon}\Phi\Big)\eta. \tag{3.15}$$

The differential equations (3.12) for p_1, q_1 then turn into

$$\dot{\eta} = \varepsilon^{-1/2} \exp\Big(-\frac{i}{\varepsilon}\Phi\Big)\, \Gamma^* \begin{pmatrix} f_1 \\ g_1 \end{pmatrix} = \varepsilon^{-1/2} P_1^* f_1 + \varepsilon^{-1/2} Q_1^* g_1$$

with the arguments $(p_0, \varepsilon^{1/2} P_1 \eta, q_0, \varepsilon^{1/2} Q_1 \eta)$ in the functions f_1, g_1. Inserting the expressions for f_1 and g_1 from (3.13), we obtain as in (2.22) and (2.23), with

$$W = -\frac{1}{2} \begin{pmatrix} L - L^T & L + L^T \\ L + L^T & L - L^T \end{pmatrix}, \tag{3.16}$$

the differential equation

$$\dot{\eta} = \exp\Big(-\frac{i}{\varepsilon}\Phi\Big) W(p_0, q_0) \exp\Big(\frac{i}{\varepsilon}\Phi\Big)\eta \tag{3.17}$$
$$+ \exp\Big(-\frac{i}{\varepsilon}\Phi\Big)\, \Gamma^* \begin{pmatrix} a(P_1\eta, P_1\eta; p_0, q_0) \\ b(P_1\eta, Q_1\eta; p_0, q_0) \end{pmatrix} \tag{3.18}$$
$$- P_1^* \Big(c(p_0, q_0) + T_1(q_0)^T \nabla U(q_0, 0) \Big) + r \tag{3.19}$$

with the remainder $r(p_0, q_0, P_1\eta, Q_1\eta) = \mathcal{O}(\varepsilon)$.

Adiabatic Invariants. For a solution with bounded energy, both $p_1(t)$ and $q_1(t)$ in (3.12) are of size $\mathcal{O}(\varepsilon^{1/2})$ and hence

$$\eta(t) = \mathcal{O}(1).$$

We now integrate both sides of the above differential equation from 0 to t. The integral of the terms in (3.19) is $\mathcal{O}(\varepsilon)$, as is seen by partial integration since $P_1^*(t)$ is oscillatory with an $\mathcal{O}(\varepsilon)$ integral and p_0, q_0 have bounded derivatives.

We now suppose that the eigenfrequencies $\omega_j(t) := \omega_j(q_0(t))$ remain separated and bounded away from 0: there is a constant $\delta > 0$ such that for any pair $\omega_j(t)$ and $\omega_k(t)$ with $j \neq k$, the lower bounds

$$|\omega_j(t) - \omega_k(t)| \geq \delta, \qquad \omega_j(t) \geq \frac{\delta}{2} \tag{3.20}$$

hold for all t under consideration. In this situation, as in Sect. XIV.2.1, the integral from 0 to t of the term (3.17) is bounded by $\mathcal{O}(\varepsilon)$, since the matrix W has zero diagonal.

It remains to study the term (3.18) with the bilinear functions a and b. This term has only oscillatory components if the following non-resonance condition is satisfied: for all j, k, l and all combinations of signs,

$$|\omega_j(t) \pm \omega_k(t) \pm \omega_l(t)| \geq \delta \tag{3.21}$$

with a positive δ independent of ε. In this case, also the integral over the term (3.18) is of size $\mathcal{O}(\varepsilon)$, and we obtain

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \leq \text{Const.} \tag{3.22}$$

If condition (3.21) is weakened to requiring that for all $j, k, l = 1, \ldots, m$,

$$\omega_j(t) \pm \omega_k(t) \pm \omega_l(t) \ \text{ has a finite number of at most simple zeros} \tag{3.23}$$

in the considered time interval, then the estimate deteriorates to (see Exercise 1)

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon^{1/2}) \qquad \text{for} \quad t \leq \text{Const.} \tag{3.24}$$

The actions

$$I_j = |\eta_j|^2 \qquad (j = 1, \ldots, m) \tag{3.25}$$

are thus adiabatic invariants:

$$I_j(t) = I_j(0) + \mathcal{O}(\varepsilon) \qquad \text{for} \quad t \leq \text{Const.} \tag{3.26}$$

in case of (3.22), and up to $\mathcal{O}(\varepsilon^{1/2})$ in case of (3.24).

The Slow System. Since the oscillatory energy equals

$$\frac{1}{2\varepsilon}\, p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon}\, q_1^T \Omega(q_0) q_1 = \sum_{j=1}^{m} I_j\, \omega_j(q_0),$$

the differential equations (3.12) for the slow variables p_0, q_0 become, up to $\mathcal{O}(\varepsilon)$,

$$\begin{aligned} \dot{p}_0 &= -\nabla_{q_0}\Big(\frac{1}{2}\, p_0^T M_0(q_0)^{-1} p_0 + U(q_0,0)\Big) - \sum_{j=1}^{m} I_j\, \nabla_{q_0}\omega_j(q_0) \\ \dot{q}_0 &= M_0(q_0)^{-1} p_0. \end{aligned} \tag{3.27}$$

Compared with the constrained system with Hamiltonian $\frac{1}{2}p^T M(q)^{-1}p + U(q)$ on the configuration manifold $\mathcal{V}$, the slow motion is thus driven by the additional potential $\sum_{j=1}^{m} I_j\, \omega_j(q_0)$ depending on the actions I_j. See also Rubin & Ungar (1957), Takens (1980), and Bornemann (1998) for different derivations and discussions of the correction potential.

Avoided Crossing of Frequencies and Takens Chaos. If the distance δ of frequencies in (3.20) becomes so small at a point $q_0(t)$ that $\delta^2 \le \varepsilon$, then there can again occur $\mathcal{O}(1)$ changes in adiabatic invariants I_j, as in the Zener example of Sect. XIV.1.1. In the present situation of solution-dependent frequencies, however, the level to which I_j jumps after the avoided crossing, depends very sensitively on the slow solution variables $q_0(t)$ through the terms $\exp(\pm\frac{i}{\varepsilon}\Phi)$ in (3.17). In turn, the slow motion of p_0, q_0 after the avoided crossing depends on the new values of I_j through (3.27). The effect is that the slow motion depends very sensitively on perturbations of the initial values in the case of an avoided crossing; see Takens (1980). The indeterminacy of the slow motion in the limit $\varepsilon \to 0$ is termed *Takens chaos* by Bornemann (1998).

XIV.3.3 Integrators in Adiabatic Variables

A long-time-step integrator for the oscillatory mechanical system with Hamiltonian (3.1) can now be obtained as follows:

Solve the slow system (3.27) in tandem with applying an adiabatic integrator (see Sect. XIV.1.2) to a simplified equation for the adiabatic variables,

$$\dot{\eta} = \exp\Big(-\frac{i}{\varepsilon}\Phi\Big) W \exp\Big(\frac{i}{\varepsilon}\Phi\Big)\eta,$$

where W is given by (3.16) with a simplified matrix L: with $v_0 = M_0(q_0)^{-1}p_0$, let

$$L(p_0,q_0) = -\Omega(q_0)^{1/2}\, \frac{d}{d\tau}\Big|_{\tau=0} Q(q_0+\tau v_0)^T\, Q(q_0)\, \Omega(q_0)^{-1/2}.$$

This matrix L captures the principal terms, coming from the matrix X_{01} in (3.8), which are responsible for a change of the adiabatic invariants due to an avoided

crossing as long as the frequency separation condition (3.20) holds with a possibly ε-dependent $\delta \gg \varepsilon$, e.g., with $\delta \sim \varepsilon^{1/2}$ where $O(1)$ changes occur in the adiabatic invariants. Because of the Takens chaos, it cannot be expected that such an integrator yields a good approximation to "the" solution, but the method can approximate an almost-solution (having a small defect in the differential equations) that passes through the avoided crossing zone, and it detects the change of adiabatic invariants. The properties of integrators of this type are currently under investigation (Lorenz & Lubich 2006).

Further we refer to Jahnke (2003, 2004b) for the construction and analysis of adiabatic integrators for mixed quantum-classical molecular dynamics, where similarly a nonlinear coupling of slow and fast, oscillatory motions occurs.

XIV.3.4 Analysis of Multiple Time-Stepping Methods

The error behaviour of the impulse and mollified impulse method applied to an oscillatory Hamiltonian system (3.1) with well-separated frequencies can be analysed in the adiabatic variables in the same way as we did in Sections XIV.2.3 and XIV.2.4 for the case of time-dependent frequencies. Analogous formulas and the same conclusions hold; essentially we need to replace the argument t by q_0 in the appearing functions. However, their behaviour in the situation of an avoided crossing with Takens chaos is presently not understood.

XIV.4 Exercises

1. Show that
$$\int_0^t \exp\Big(\frac{i}{\varepsilon}\,\phi(s)\Big)ds = \mathcal{O}(\varepsilon^{1/(m+1)})$$
if $\lambda := \dot\phi$ has finitely many zeros of order at most m in the interval $[0,t]$.
Hint: Use the *method of stationary phase*; see, e.g., Olver (1974) or van der Corput (1934).
2. Show that the adiabatic variables $\eta(t)$ of (1.4) remain approximately constant also in the following cases of non-separated eigenvalues:
(a) a multiple eigenvalue $\lambda_j(t)$ of constant multiplicity m for all t and the orthogonal basis $v_{j,1}(t),\dots,v_{j,m}(t)$ of the corresponding eigenspace chosen such that the derivatives $\dot v_{j,l}(t)$ are orthogonal to the eigenspace for all t;
(b) a crossing of eigenvalues, $\lambda_j(t_*) = \lambda_k(t_*)$ with $\dot\lambda_j(t_*) \neq \dot\lambda_k(t_*)$, for which the eigenvectors are smooth functions of t in a neighbourhood of t_*; see also Born & Fock (1928) for crossings where $\lambda_j - \lambda_k$ can have zeros of higher multiplicity.
3. Let the differential equation (1.1) with smooth skew-hermitian $Z(t)$ be transformed locally over $[t_0, t_0+h]$ to $z(t) = \exp(-\frac{t}{\varepsilon}Z_*)y(t)$, so that

$$\dot{z} = \frac{1}{\varepsilon} \exp(-\frac{t}{\varepsilon} Z_*) \big(Z(t) - Z_*\big) \exp(\frac{t}{\varepsilon} Z_*)\, z$$

with $Z_* = Z(t_0 + h/2)$. Consider the averaged midpoint rule

$$z_1 = z_0 + \frac{1}{\varepsilon} \int_0^h \exp(-\frac{s}{\varepsilon} Z_*) \big(\widetilde{Z}(s) - Z_*\big) \exp(\frac{s}{\varepsilon} Z_*)\, ds\; \frac{1}{2}(z_0 + z_1), \quad (4.1)$$

where $\widetilde{Z}(t)$ is the quadratic interpolation polynomial through $Z(t_0)$, Z_*, $Z(t_1)$. Show that the local error $z_1 - z(t_1)$ is of size $\mathcal{O}(h^4/\varepsilon^2)$, which is $\mathcal{O}(h^2)$ only for $h = \mathcal{O}(\varepsilon)$. Explain why the error bound cannot be improved to $\mathcal{O}(h^2)$ for $h = \mathcal{O}(\varepsilon^\alpha)$ with $\alpha < 1$.
Hint: See the proofs of Theorems 2.1(i) and 3.1 in Hochbruck & Lubich (1999b), cf. also Iserles (2004).

4. In the situation of the previous exercise, let U be a unitary matrix of eigenvectors of Z_*, and let $\widetilde{D}(t)$ be the diagonal matrix containing the diagonal entries of $U^*\big(\widetilde{Z}(t) - Z_*\big)U$. Find a modification of the above averaged midpoint rule by terms that use only $\widetilde{D}(t)$, such that the local error is $\mathcal{O}(h^2)$ for $h \le \varepsilon^{3/4}$ if the eigenvalues of Z_* are all separated by a distance δ independent of ε.
5. Compare the error behaviour of the averaged midpoint rules (1.12) and (4.1) near the avoided crossing of the eigenvalues in the Zener matrix (1.9).
6. Formulate symmetric modifications of the adiabatic integrators (1.12) and (1.13) that use function evaluations at the grid points t_n and t_{n+1} instead of $t_{n+1/2}$.
7. Consider the differential equation $\dot{y} = f(y) + g(t)$ with a smooth function $f(y)$ and a function $g(t) = \mathcal{O}(1)$ with $\dot{g}(t) = \mathcal{O}(\delta^{-1})$ with respect to a small parameter δ. For the modified midpoint rule

$$y_1 = y_0 + hf\Big(\frac{y_0 + y_1}{2}\Big) + \int_{t_0}^{t_1} g(t)\, dt,$$

show that the local error satisfies $y_1 - y(t_1) = \mathcal{O}(h^3/\delta)$.
8. Write the Hamiltonian system (XIII.9.2) in adiabatic variables and relate this to the first terms of the modulated Fourier expansion.
9. Compare the impulse method of Sect. XIV.2.3 with the method based on the splitting

$$H = \Big(\frac{1}{2}\, p^T M(t)^{-1} p + \frac{1}{2\varepsilon^2}\, q^T A(t) q\Big) + \Big(U(q,t) + E\Big).$$

10. Show that Theorem 2.6 remains valid for the choice $\mathcal{S}(t) = 0$ in (2.37). This corresponds to the projection to the constraint manifold in the mollified impulse method as proposed by Izaguirre, Reich & Skeel (1999).

Chapter XV. Dynamics of Multistep Methods

Multistep methods are the basis of important codes for nonstiff differential equations (Adams methods) and for stiff problems (BDF methods). We study here their applicability to long-time integrations of Hamiltonian or reversible systems.

This chapter starts with numerical experiments which illustrate that the long-time behaviour of classical multistep methods is in general disappointing. They either behave as non-symplectic and non-symmetric one-step methods, or they exhibit undesired instabilities (parasitic solutions). Certain multistep methods for second order equations or partitioned multistep methods, however, have a much better long-time behaviour. They are promising methods, because in a constant step size mode they can be easily implemented, and high order can be obtained with one function evaluation per step. We characterize such methods by studying their underlying one-step method, their symplecticity, their conservation properties, as well as their long-term stability.

XV.1 Numerical Methods and Experiments

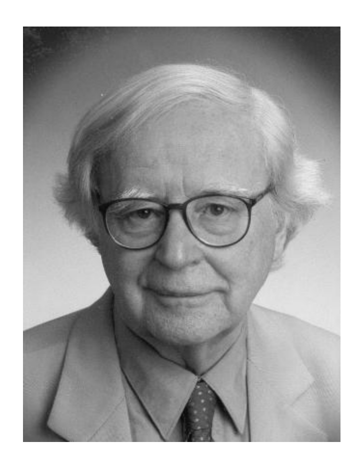
We present the numerical methods treated in this chapter, and in numerical experiments we look at their behaviour on Hamiltonian systems.

XV.1.1 Linear Multistep Methods

For first order systems of differential equations $\dot y = f(y)$, linear multistep methods are defined by the formula

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(y_{n+j}), \tag{1.1}$$

where α_j, β_j are real parameters, $\alpha_k \neq 0$, and $|\alpha_0| + |\beta_0| > 0$. For an application of this formula we need a starting procedure which, in addition to an initial value $y(t_0) = y_0$, provides approximations $y_1, \ldots, y_{k-1}$ to $y(t_0{+}h), \ldots, y(t_0{+}(k{-}1)h)$. The approximations y_n to $y(t_0 + nh)$ for $n \geq k$ can then be computed recursively from (1.1). In the case $\beta_k = 0$ we have an explicit method, otherwise it is implicit and the numerical solution y_{n+k} has to be computed iteratively.


Germund Dahlquist[1]

Since the fundamental work of Dahlquist (1956) it is common to denote the generating polynomials of the coefficients by

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

For the classical theory of multistep methods we refer the reader to Chap. III of Hairer, Nørsett & Wanner (1993). We just recall some important definitions.

Order. A multistep method has order r if, when applied with exact starting values to the problem $\dot{y} = t^q$ $(0 \le q \le r)$, it integrates the problem without error. This is equivalent to the requirement that

$$\rho(e^h) - h\sigma(e^h) = \mathcal{O}(h^{r+1}) \qquad \text{for} \quad h \to 0. \tag{1.2}$$

Stability. Method (1.1) is stable if, when applied to $\dot{y} = 0$, it yields for all $y_0, \ldots, y_{k-1}$ a bounded numerical solution. This is equivalent to the requirement that the polynomial $\rho(\zeta)$ satisfies the root condition, i.e., all roots of $\rho(\zeta) = 0$ satisfy $|\zeta| \le 1$, and those on the unit circle are simple roots. The method is called *strictly stable*, if all roots are inside the unit circle with the exception of $\zeta = 1$.

Convergence. If a multistep method is stable and of order $r \ge 1$, it is convergent of order r for all sufficiently smooth problems. This means that, assuming starting approximations with an error bounded by $\mathcal{O}(h^r)$, the global error satisfies $y_n - y(t_0 + nh) = \mathcal{O}(h^r)$ on compact intervals $nh \le T$.

Symmetry. If the coefficients of a multistep formula (1.1) satisfy

$$\alpha_{k-j} = -\alpha_j, \qquad \beta_{k-j} = \beta_j \qquad \text{for all } j, \tag{1.3}$$

then the method is called symmetric. Condition (1.3) implies that for every zero ζ of $\rho(\zeta)$ also its inverse ζ^{-1} is a zero. Hence, for stable symmetric methods all zeros of $\rho(\zeta)$ are simple and lie on the unit circle.

Example 1.1. We consider the pendulum equation (I.1.13), and we apply the following multistep methods: the 2-step explicit Adams method

$$y_{n+2} = y_{n+1} + h\Big(\frac{3}{2} f_{n+1} - \frac{1}{2} f_n\Big), \tag{1.4}$$

the 2-step backward differentiation formula (BDF)

$$\frac{3}{2} y_{n+2} - 2y_{n+1} + \frac{1}{2} y_n = h f_{n+2}, \tag{1.5}$$

and the (2-step) symmetric explicit midpoint rule

[1] Germund Dahlquist, born: 16 January 1925 in Uppsala (Sweden), died: 8 February 2005.

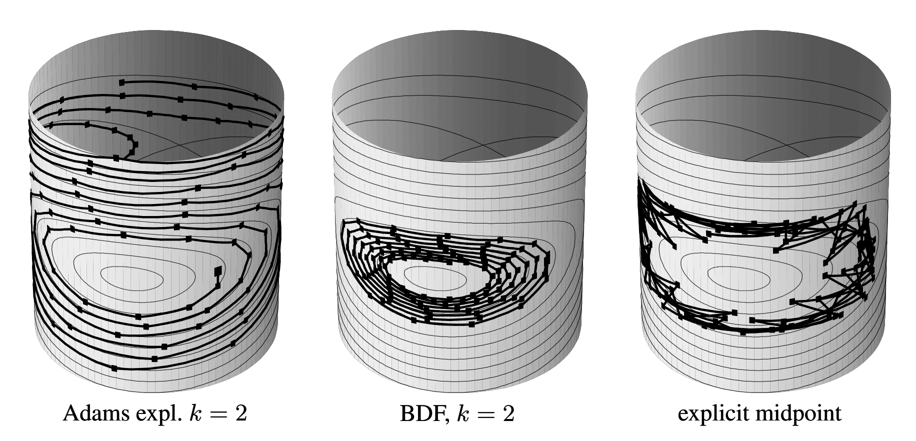


Fig. 1.1. Solutions of the pendulum problem (I.1.13); explicit Adams with step size $h = 0.5$, initial value $(p_0, q_0) = (0, 0.7)$; BDF with step size $h = 0.5$, initial value $(p_0, q_0) = (0, 0.95)$; explicit midpoint rule with $h = 0.4$ and initial value $(p_0, q_0) = (1.1, 0)$

$$y_{n+2} = y_n + 2hf_{n+1}. \tag{1.6}$$

For all methods we take $y_1 = y_0 + hf_0$ as the approximation for $y(t_0 + h)$. The results of the first 108 steps are shown in Fig. 1.1. We observe that the first two methods, as expected, behave similarly as the explicit and implicit Euler method (the numerical solution spirals either outwards or inwards). This will be rigorously explained in Sect. XV.2.1 below. However, as might not be expected, the symmetric method (1.6) does not behave like the implicit midpoint rule (cf. Fig. I.1.4), it shows undesired increasing oscillations (parasitic solutions).

After this negative experience with classical multistep methods, the obvious question is: are there multistep methods which have a long-time behaviour that is comparable to symplectic and/or symmetric one-step methods?

XV.1.2 Multistep Methods for Second Order Equations

Many important Hamiltonian systems are second order differential equations

$$\ddot{y} = f(y), \tag{1.7}$$

where the force f is independent of the velocity $\dot{y}$. Introducing the new variable $v = \dot{y}$, we obtain the system $\dot{y} = v$, $\dot{v} = f(y)$ of first order equations. If we apply a multistep method (1.1) with generating polynomials $\rho^*(\zeta) = \sum_{j=0}^{k^*} \alpha_j^* \zeta^j$ and $\sigma^*(\zeta) = \sum_{j=0}^{k^*} \beta_j^* \zeta^j$ to this system, we get

$$\sum_{j=0}^{k^*} \alpha_j^* y_{n+j} = h \sum_{j=0}^{k^*} \beta_j^* v_{n+j}, \qquad \sum_{j=0}^{k^*} \alpha_j^* v_{n+j} = h \sum_{j=0}^{k^*} \beta_j^* f(y_{n+j}).$$

An elimination of the v-variables then yields

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h^2 \sum_{j=0}^{k} \beta_j f(y_{n+j}), \tag{1.8}$$

where $k = 2k^*$, $\rho(\zeta) = \rho^*(\zeta)^2$ and $\sigma(\zeta) = \sigma^*(\zeta)^2$. We consider here methods (1.8) which do not necessarily originate from a multistep method for first order equations, and we denote the generating polynomials of the coefficients α_j and β_j again by $\rho(\zeta)$ and $\sigma(\zeta)$. From the classical theory (see Sect. III.10 of Hairer, Nørsett & Wanner 1993) we recall the following definitions and results.

Order. A method (1.8) has order r if its generating polynomials satisfy

$$\rho(e^h) - h^2\sigma(e^h) = \mathcal{O}(h^{r+2}) \qquad \text{for} \quad h \to 0. \tag{1.9}$$

Stability. Method (1.8) is stable if all zeros of the polynomial $\rho(\zeta)$ satisfy $|\zeta| \le 1$, and those on the unit circle are at most double zeros. Observe that for methods originating from (1.1) all zeros are double. The method is called *strictly stable*, if all zeros are inside the unit circle with the exception of $\zeta = 1$.

Convergence. If a multistep method (1.8) is stable, of order $r \ge 1$ and if the starting values are accurate enough, the global error satisfies $y_n - y(t_0 + nh) = \mathcal{O}(h^r)$ on compact intervals $nh \le T$.

Symmetry. If the coefficients of (1.8) satisfy

$$\alpha_{k-j} = \alpha_j, \qquad \beta_{k-j} = \beta_j \qquad \text{for all } j, \tag{1.10}$$

then the method is symmetric. Again, for every zero ζ of $\rho(\zeta)$ the value ζ^{-1} is also a zero. Hence, stable symmetric methods have all zeros of $\rho(\zeta)$ on the unit circle and they are at most of multiplicity two.

Dahlquist (1956) noticed that double zeros of $\rho(\zeta)$ on the unit circle can lead to an exponential error growth. Lambert & Watson (1976) analyzed in detail the application of (1.8) to the linear test equation $\ddot{y} = -\omega^2 y$. They found that with symmetric methods for which $\rho(\zeta)$ does not have double roots on the unit circle other than $\zeta = 1$, the numerical solution remains close to a periodic orbit (for sufficiently small step sizes). For example, the Störmer–Verlet method $y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$ satisfies this property for $0 < h\omega < 2$ (see Sect. I.5.2). The study of the long-time behaviour of symmetric methods (1.8) was then put forward by the article of Quinlan & Tremaine (1990), where an excellent performance for simulations of the outer solar system is reported.

Example 1.2. We consider the Kepler problem (I.2.2) with initial values (I.2.11) and eccentricity $e = 0.2$. We apply the following three methods with constant step size $h = 0.01$ on the interval of length $2\pi \cdot 10^5$ (i.e., 10^5 periods):

(A) $\quad y_{n+4} - 2y_{n+3} + y_{n+2} = h^2\Big(\frac{7}{6}f_{n+3} - \frac{5}{12}f_{n+2} + \frac{1}{3}f_{n+1} - \frac{1}{12}f_n\Big)$

(B) $\quad y_{n+4} - 2y_{n+2} + y_n = h^2\Big(\frac{4}{3}f_{n+3} + \frac{4}{3}f_{n+2} + \frac{4}{3}f_{n+1}\Big)$

(C) $\quad y_{n+4} - 2y_{n+3} + 2y_{n+2} - 2y_{n+1} + y_n = h^2\Big(\frac{7}{6}f_{n+3} - \frac{1}{3}f_{n+2} + \frac{7}{6}f_{n+1}\Big).$

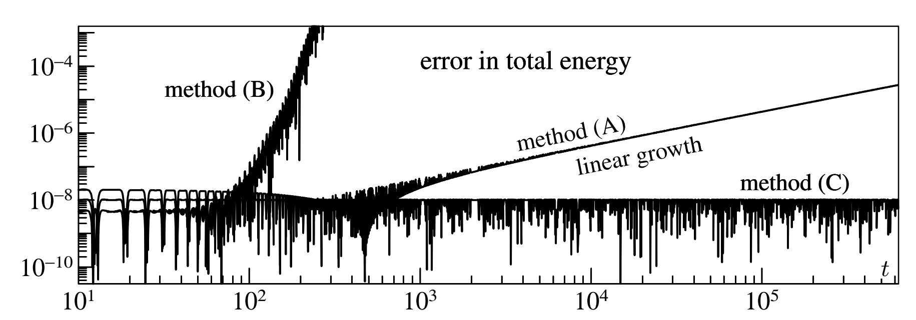


Fig. 1.2. Error in the total energy for the three linear multistep methods of Example 1.2 applied to the Kepler problem with $e = 0.2$

All three methods are of order $r = 4$; method (A) is strictly stable, whereas methods (B) and (C) are symmetric. For method (B) the ρ-polynomial has a double root at $\zeta = -1$, for method (C) it does not have double roots other than 1. Starting values y_1, y_2, and y_3 are computed very accurately with a high-order Runge–Kutta method.

The error in the total energy is plotted for all three methods in Fig. 1.2. On the first 10 periods, all methods behave similarly and no error growth is observed. Beyond this interval, method (A) shows a linear error growth (as it is the case for non-symplectic and non-symmetric one-step methods), method (B) has an exponential error growth, and for method (C) the error remains bounded of size $\mathcal{O}(h^4)$ on the whole interval of integration. One of the aims of this chapter is to explain the excellent long-time behaviour of method (C).

Stabilized Version of (1.8). Due to the double zeros (of modulus one) of the characteristic polynomial of the difference equation $\sum_j \alpha_j y_{n+j} = 0$, we have an undesired propagation of rounding errors (especially for long-time integrations). To overcome this difficulty, we split the characteristic polynomial $\rho(\zeta)$ into

$$\rho(\zeta) = \rho_A(\zeta) \cdot \rho_B(\zeta), \tag{1.11}$$

such that each polynomial

$$\rho_A(\zeta) = \sum_{j=0}^{k_A} \alpha_j^{(A)} \zeta^j, \qquad \rho_B(\zeta) = \sum_{j=0}^{k_B} \alpha_j^{(B)} \zeta^j$$

has only simple roots of modulus one. Introducing the new variable $hv_n := \sum_j \alpha_j^{(A)} y_{n+j}$, the recurrence relation (1.8) becomes equivalent to

$$\sum_{j=0}^{k_A} \alpha_j^{(A)} y_{n+j} = hv_n, \qquad \sum_{j=0}^{k_B} \alpha_j^{(B)} v_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}. \tag{1.12}$$

This formula, which for the Störmer–Verlet scheme corresponds to the one-step formulation (I.1.17), is much better suited for an implementation. If the splitting is such that $\rho_A'(1) = 1$, the discretization (1.12) is consistent with the first order partitioned system $\dot{y} = v, \dot{v} = f(y)$.

XV.1.3 Partitioned Multistep Methods

Motivated by the stabilized version (1.12) of multistep methods for second order equations, let us consider general partitioned systems of differential equations

$$\dot{y} = f(y, v), \qquad \dot{v} = g(y, v), \tag{1.13}$$

where, needless to say, y and v may be vectors. The idea is to apply different multistep methods to different components. We thus get

$$\sum_{j=0}^{k} \alpha_j^{(A)} y_{n+j} = h \sum_{j=0}^{k} \beta_j^{(A)} f_{n+j}, \qquad \sum_{j=0}^{k} \alpha_j^{(B)} v_{n+j} = h \sum_{j=0}^{k} \beta_j^{(B)} g_{n+j}, \tag{1.14}$$

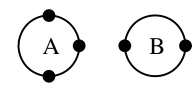
where $f_n = f(y_n, v_n)$ and $g_n = g(y_n, v_n)$. We can take the same k for both methods without loss of generality, if we abandon the assumption $|\alpha_0| + |\beta_0| > 0$.

Such a method is of order r, if both methods are of order r. It is stable (strictly stable, symmetric, ...), if both methods are stable (strictly stable, symmetric, ...).

Example 1.3. For our next experiment we use the symmetric methods

$$\begin{aligned} \text{(A)}: &\quad y_{n+3} - y_{n+2} + y_{n+1} - y_n = h(f_{n+2} + f_{n+1}) \\ \text{(B)}: &\quad v_{n+3} - v_{n+1} = 2h g_{n+2}. \end{aligned} \tag{1.15}$$

Both methods are of order 2, and their ρ-polynomials $\rho_A(\zeta) = (\zeta - 1)(\zeta^2 + 1)$ and $\rho_B(\zeta) = (\zeta - 1)(\zeta + 1)$ do not have common zeros with the exception of $\zeta = 1$.



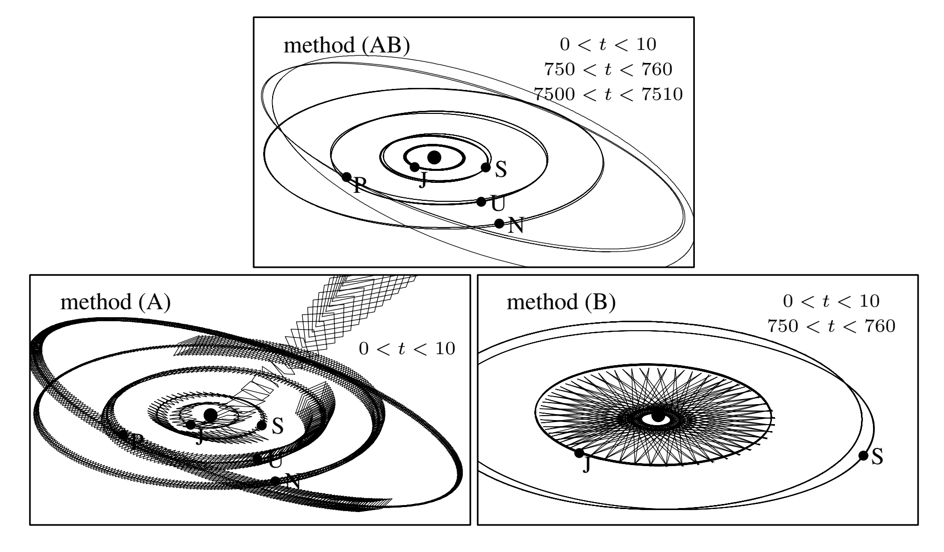


Fig. 1.3. Three versions of the methods (1.15) applied with step size $h = 50$ (days) to the outer solar system. For method (B) only the numerical orbits of Jupiter and Saturn are plotted. The time intervals are given in units of 10 000 days

We choose the outer solar system with the data as described in Sect. I.2.4, and we apply the methods in three versions: (i) as partitioned method (AB), where the positions are treated by method (A) and the velocities by method (B); (ii) method (A) is applied to all components; (iii) method (B) is applied to all components. The numerical results are shown in Fig. 1.3. Whereas the individual methods show instabilities on rather short time intervals, the partitioned method gives a correct picture even with a large step size $h = 50$.

XV.2 The Underlying One-Step Method

Much insight into the long-time behaviour of multistep methods can be gained by relating their numerical solution to one-step methods. This then allows for an application of the considerations of the preceding sections.

XV.2.1 Strictly Stable Multistep methods

It was a surprising result when Kirchgraber (1986) proved that strictly stable multistep methods are essentially equivalent to one-step methods. Although this one-step method is "quite exotic" (Eirola & Nevanlinna 1988), it is the key for a better understanding of the dynamics of strictly stable methods.

Theorem 2.1 (Kirchgraber 1986). *Consider a strictly stable linear multistep method (1.1) applied with a sufficiently small step size h. Then, there exists a one-step method Φ_h such that for starting approximations computed by $y_j = \Phi_h^j(y_0)$, $j = 1, \dots, k-1$, the numerical solution of (1.1) is identical to that obtained by the one-step method, i.e., $y_{n+1} = \Phi_h(y_n)$ for all $n \geq 0$.*

Proof. The idea is to reformulate the multistep method (1.1) in such a way that the Invariant Manifold Theorem of Sect. XII.3 can be applied. To keep the notation as simple as possible, let us consider the case $k = 3$.

We write the method in the form

$$\begin{pmatrix} y_{n+3} \\ y_{n+2} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{n+2} \\ y_{n+1} \\ y_n \end{pmatrix} + h \begin{pmatrix} F_h(y_n, y_{n+1}, y_{n+2}) \\ 0 \\ 0 \end{pmatrix} \tag{2.1}$$

with $a_i = \alpha_i / \alpha_k$, and we transform the appearing matrix A to Jordan canonical form $J = T^{-1}AT$. We thus get

$$Z_{n+1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & d_{11} & d_{12} \\ 0 & d_{21} & d_{22} \end{pmatrix} Z_n + hG(Z_n), \qquad Z_n = T^{-1} \begin{pmatrix} y_{n+2} \\ y_{n+1} \\ y_n \end{pmatrix}. \tag{2.2}$$

Since the method is strictly stable, 1 is a simple eigenvalue of A, and all other eigenvalues are less than 1 in modulus. Consequently, the matrix $D = (d_{ij})$ satisfies

$\|D\| < 1$ in a suitable norm. Partitioning $Z_n = (\xi_n, \eta_n)^T$ into its first component ξ_n and the rest (collected in η_n), we see that (2.2) is of the form (XII.3.1) with L_{xx}, L_{xy}, L_{yx} of size $\mathcal{O}(h)$, and $L_{yy} = \|D\| < 1$. Theorem XII.3.1 thus yields the existence of a function $\eta = s(\xi)$ such that the manifolds

$$\mathcal{N}_h = \left\{ \begin{pmatrix} \xi \\ s(\xi) \end{pmatrix} ; \xi \in \mathbb{R}^d \right\} \qquad \text{and} \qquad \mathcal{M}_h = \left\{ T \begin{pmatrix} \xi \\ s(\xi) \end{pmatrix} ; \xi \in \mathbb{R}^d \right\}$$

are invariant under the mappings (2.2) and (2.1), respectively. The function $s(\xi)$ is Lipschitz continuous with constant $\lambda = \mathcal{O}(h)$.

Since the first column of T, which is the eigenvector corresponding to the eigenvalue 1 of A, is given by $(1, 1, 1)^T$, the last component of $T\binom{\xi}{s(\xi)}$ satisfies $y = \xi + g(\xi)$ where $g(\xi)$ is Lipschitz continuous with constant $\mathcal{O}(h)$. By the Banach fixed-point theorem this equation has a unique solution $\xi = r(y)$. Consequently, the manifold $\mathcal{M}_h$ can be parametrized in terms of y as

$$\mathcal{M}_h = \left\{ (\Psi_h(y), \Phi_h(y), y)^T ; y \in \mathbb{R}^d \right\}.$$

Its invariance under (2.1) implies that

$$\begin{pmatrix} \Psi_h(\widehat{y}) \\ \Phi_h(\widehat{y}) \\ \widehat{y} \end{pmatrix} = \begin{pmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Psi_h(y) \\ \Phi_h(y) \\ y \end{pmatrix} + h \begin{pmatrix} F_h(y, \Phi_h(y), \Psi_h(y)) \\ 0 \\ 0 \end{pmatrix}$$

and consequently $\widehat{y} = \Phi_h(y)$ and $\Phi_h(\widehat{y}) = \Psi_h(y)$, so that $\Psi_h(y) = \Phi_h^2(y)$. This holds for all y, and thus proves the statement of the theorem. □

Example 2.2. For a scalar linear problem $\dot{y} = \lambda y$, the application of a multistep method yields a difference equation with characteristic polynomial $\rho(\zeta) - h\lambda\sigma(\zeta)$. Denoting its zeros by $\zeta_1(h\lambda), \ldots, \zeta_k(h\lambda)$, where $\zeta_1(0) = 1$ and $|\zeta_j(0)| < 1$ for $j \geq 2$, the numerical solution can be written as (assuming distinct $\zeta_j(h\lambda)$)

$$y_n = c_1\zeta_1^n(h\lambda) + c_2\zeta_2^n(h\lambda) + \ldots + c_k\zeta_k^n(h\lambda).$$

The coefficients $c_1, \ldots, c_k$ depend on $h\lambda$ and are determined by the starting approximations $y_0, \ldots, y_{k-1}$. In this situation the underlying one-step method is the mapping $y_0 \mapsto \zeta_1(h\lambda)y_0$. Observe that $\zeta_1(z)$ is in general not a rational function as we are used to with Runge–Kutta methods.

Remark 2.3 (Asymptotic Phase). For arbitrary $y_0, y_1, \ldots, y_{k-1}$ close to the exact solution, there exists y_0^* such that the multistep solution $\{y_n\}$ and the one-step solution $\{y_n^*\}$, given by $y_{n+1}^* = \Phi_h(y_n^*)$, approach exponentially fast, i.e.,

$$\|y_n - y_n^*\| \leq Const \cdot \rho^n \qquad \text{for all } n \geq 0 \tag{2.3}$$

with some ρ satisfying $0 < \rho < 1$ (see Exercise XII.3). This is due to the attractivity of the invariant manifold $\mathcal{M}_h$. A proof is given in Stoffer (1993), and it is based on techniques of Nipp & Stoffer (1992). This result explains why strictly stable linear multistep methods have the same long-time behaviour as one-step methods.

In the context of "geometric numerical integration" we are mainly interested in symplectic and/or symmetric methods which, for linear problems, are characterized by the condition $\zeta_1(-z)\zeta_1(z) \equiv 1$ (see Sect. VI.4.2). This, however, is only possible for symmetric multistep methods (Exercise 1) which cannot be strictly stable.

XV.2.2 Formal Analysis for Weakly Stable Methods

The proof and the statement of Theorem 2.1 break down as soon as at least one root of $\rho(\zeta)$, different from 1, has modulus one. Moreover, Example 2.2 shows that we cannot expect a property like (2.3) with $\rho < 1$. All we can hope for is to find an underlying one-step method as a formal series in h. Surprisingly, this provides a lot of insight into the long-time dynamics of weakly stable multistep methods.

Theorem 2.4. *Consider a linear multistep method (1.1), and assume that* $\zeta = 1$ *is a single root of* $\rho(\zeta) = 0$. *Then there exists a unique formal expansion*

$$\Phi_h(y) = y + hd_1(y) + h^2 d_2(y) + \dots \tag{2.4}$$

such that

$$\sum_{j=0}^{k} \alpha_j \Phi_h^j(y) \;=\; h \sum_{j=0}^{k} \beta_j f\big(\Phi_h^j(y)\big),$$

where identity is understood in the sense of formal power series in h.

If the multistep method is of order r*, then also the underlying one-step method is of order* r*, i.e.,* $\Phi_h(y) - \varphi_h(y) = \mathcal{O}(h^{r+1})$.

The formal series for $\Phi_h(y)$ is called "step-transition operator" in the Chinese literature (see e.g., Feng (1995), page 274). We call it "underlying one-step method". Notice that this theorem does not require any stability assumption.

Proof. Expanding $\Phi_h^j(y)$ and $f\big(\Phi_h^j(y)\big)$ into powers of h, a comparison of the coefficients yields

$$\begin{aligned}
\rho'(1)\, d_1(y) &= \sigma(1)\, f(y)\\
\rho'(1)\, d_2(y) &= -\frac{\rho''(1)}{2}\, d_1'(y) d_1(y) \;+\; \sigma'(1)\, f'(y) d_1(y)\\
\rho'(1)\, d_j(y) &= \dots ,
\end{aligned} \tag{2.5}$$

where the three dots represent known functions depending on derivatives of $f(y)$ and on $d_i(y)$ with $i < j$. Since $\rho'(1) \neq 0$ by assumption, unique coefficient functions $d_j(y)$ are obtained recursively. The statement on the order follows from the fact that the exact flow $\varphi_h(y)$ has a defect $\mathcal{O}(h^{r+1})$ in the multistep formula. □

The computation of the previous proof shows that the series (2.4) is a B-series. This follows rigorously from the results of Sect. III.1.4. Whereas the B-series representation of Runge–Kutta methods converges for sufficiently small h, this is in general not the case for (2.4); see the next example.

Example 2.5. Consider a consistent two-step method

$$\alpha_2 y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h\big(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n\big),$$

and apply it to the simple system $\dot y = f(t),\ \dot t = 1$. The y-component of the underlying one-step method then takes the form

$$\Phi_h(t_0, y_0) = y_0 + \sum_{j\ge 1} h^j \, a_j \, f^{(j-1)}(t_0). \tag{2.6}$$

Putting $f(t) = e^t$ yields

$$A(\zeta) = \sum_{j\ge 1} a_j \, \zeta^{j-1} = \frac{\beta_2 e^{2\zeta} + \beta_1 e^{\zeta} + \beta_0}{\alpha_2(1 + e^{\zeta}) + \alpha_1}.$$

for the generating function of the coefficients a_j. Since this function has finite poles, the radius of convergence of $A(\zeta)$ is finite. Therefore, the radius of convergence of the series (2.6) has to be zero as soon as $f^{(j)}(t_0)$ behaves like $j!\,\mu\,\kappa^j$ (this is typically the case for analytic functions). Independent of the fact whether the method is strictly stable or not, the series (2.6) usually does not converge.

Both, Theorem 2.1 and Theorem 2.4, extend in a straightforward manner to partitioned multistep methods (1.14). To get analogous results for multistep methods (1.8) for second order differential equations, one has to introduce an approximation for the velocity $v = \dot y$. This will be explained in more detail in Sect. XV.3 below.

XV.3 Backward Error Analysis

The backward error analysis for multistep methods (Hairer 1999) is presented in two steps:

- for "smooth" numerical solutions (obtained by the underlying one-step method);
- for the general case.

The idealized situation of no parasitic terms gives already much insight into conservation properties of the method (see Sect. XV.4). The study of the general case is, however, necessary for getting estimates for the parasitic solutions (Sect. XV.5), so that rigorous statements on the long-time behaviour are possible.

XV.3.1 Modified Equation for Smooth Numerical Solutions

The formal backward error analysis of Chap. IX could be directly applied to the underlying one-step method of Sect. XV.2.2. However, due to the non-convergence of the series for $\Phi_h(y)$, difficulties may arise as soon as rigorous estimates are desired. We prefer to derive the modified differential equation directly from the multistep formula and thus avoid the use of the underlying one-step method.

Theorem 3.1. *Consider a linear multistep method (1.1), and assume that* $\rho(1)=0$ *and* $\rho'(1)=\sigma(1)\neq 0$. *Then there exist unique* h*-independent functions* $f_j(y)$ *such that, for every truncation index* N, *every solution of*

$$\dot y = f(y) + hf_2(y) + h^2 f_3(y) + \ldots + h^{N-1} f_N(y) \tag{3.1}$$

satisfies

$$\sum_{j=0}^{k} \alpha_j y(t+jh) \;=\; h\sum_{j=0}^{k} \beta_j f\bigl(y(t+jh)\bigr) \;+\; \mathcal{O}(h^{N+1}). \tag{3.2}$$

If the multistep method is of order r, *then* $f_j(y)=0$ *for* $2\le j\le r$. *If the method is symmetric, then* $f_j(y)=0$ *for all even* j, *so that the modified equation (3.1) has an expansion in even powers of* h.

Proof. Using the Lie derivative $(D_i g)(y) = g'(y) f_i(y)$ (with $f_1(y)=f(y)$) and $D = D_1 + hD_2 + h^2 D_3 + \ldots$, the solution of (3.1) with initial value $y(t)=y$ satisfies $y(t+jh) = e^{jhD} y + \mathcal{O}(h^{N+1})$ and $f(y(t+jh)) = e^{jhD} f(y) + \mathcal{O}(h^{N+1})$ (by Taylor expansion). We thus have

$$\rho(e^{hD}) y \;=\; h\sigma(e^{hD}) f(y) \;+\; \mathcal{O}(h^{N+1}). \tag{3.3}$$

With the expansion $x\sigma(e^x)/\rho(e^x) = 1 + \mu_1 x + \mu_2 x^2 + \ldots$ this becomes

$$\dot y = (1 + \mu_1 hD + \mu_2 h^2 D^2 + \ldots) f(y) + \mathcal{O}(h^N). \tag{3.4}$$

A comparison with (3.1) yields $f_1(y) = f(y)$, and

$$f_j(y) \;=\; \sum_{l\ge 1} \mu_l \sum_{j_1+\ldots+j_l=j-1} \Bigl(D_{j_1}\ldots D_{j_l} f\Bigr)(y) \tag{3.5}$$

for $j\ge 2$, which uniquely defines the functions $f_j(y)$ in a recursive manner. □

Lemma 3.2. *If* $f(y)$ *is analytic and bounded by* M *in* $B_R(y_0)$, *then we have*

$$\bigl\|f_j(y)\bigr\| \le \mu\, M \left(\frac{\eta\, M\, j}{R}\right)^{j-1} \qquad \textit{for} \qquad \|y-y_0\| \le R/2, \tag{3.6}$$

where μ *and* η *depend only on the coefficients* α_j, β_j *of the multistep method.*

Proof. The estimate (3.6) is obtained as in the proof of Theorem IX.7.5. We just sketch the main idea in the notation used there. With $\delta = R/(2(J-1))$ we have $\|f_j\|_j \le \delta b_j$, where the generating function $b(\zeta) = \sum_{j\ge 1} b_j \zeta^j$ of the b_j satisfies

$$b(\zeta) \;=\; \frac{M\zeta}{\delta}\Bigl(1 + \sum_{l\ge 1} |\mu_l|\, b(\zeta)^l\Bigr).$$

By the implicit function theorem, $b(\zeta)$ is analytic and bounded in a disc of radius $c\delta/M$ centred at the origin (c is a positive constant depending only on the coefficients of the multistep method). The estimate (3.6) then follows from Cauchy's inequalities as in the proof of Theorem IX.7.5. □

It is remarkable that, although the Taylor series of the underlying one-step method generally diverges, the coefficient functions of the modified differential equation satisfy the same estimate as for Runge–Kutta methods. This enables us to prove an analogue of Theorem IX.7.6 which, for one-step methods, is the main ingredient for exponentially small error estimates. One can prove that for suitably chosen $N = N(h)$ and for $h \leq h_0/4$ with $h_0 = R/(e\eta M)$, the solution of (3.1) satisfies

$$\Big\| \sum_{j=0}^{k} \alpha_j y(t+jh) - h \sum_{j=0}^{k} \beta_j f\big(y(t+jh)\big) \Big\| \leq h\gamma M e^{-h_0/h},$$

where γ depends only on the multistep formula. The proof of this statement is similar to that of Theorem IX.7.6. We skip details and refer to Hairer (1999).

For strictly stable multistep methods, Theorem 2.1 together with the Invariant Manifold Theorem XII.3.2 thus imply that the underlying one-step method is exponentially close to the exact solution of the truncated modified equation. The parasitic solution terms are rapidly damped out by the property (2.3) of asymptotic phase. The same conclusions as for one-step methods can therefore be drawn.

For symmetric methods the situation is not so simple. One has to study the parasitic solution components to get information on the long-time behaviour of the numerical solution of (1.1). The basic techniques will be prepared in Sect. XV.3.2.

Partitioned Multistep Methods. The extension of the modified differential equation to methods (1.14) is straightforward. There exist functions $f_j(y,v)$ and $g_j(y,v)$ such that the exact solution of

$$\begin{aligned} \dot y &= f(y,v) + hf_2(y,v) + \ldots + h^{N-1} f_N(y,v) \\ \dot v &= g(y,v) + hg_2(y,v) + \ldots + h^{N-1} g_N(y,v) \end{aligned} \tag{3.7}$$

satisfies the multistep formula (1.14) up to a defect of size $\mathcal{O}(h^{N+1})$. The coefficient functions can be computed by comparing (3.7) to

$$\begin{aligned} \dot y &= (1 + \mu_1^{(A)} hD + \mu_2^{(A)} h^2 D^2 + \ldots) f(y,v) + \mathcal{O}(h^N) \\ \dot v &= (1 + \mu_1^{(B)} hD + \mu_2^{(B)} h^2 D^2 + \ldots) g(y,v) + \mathcal{O}(h^N), \end{aligned} \tag{3.8}$$

where the real numbers $\mu_j^{(A)}$ and $\mu_j^{(B)}$ are given by $x\sigma^{(A)}(e^x)/\rho^{(A)}(e^x) = 1 + \mu_1^{(A)} x + \mu_2^{(A)} x^2 + \ldots$ and by $x\sigma^{(B)}(e^x)/\rho^{(B)}(e^x) = 1 + \mu_1^{(B)} x + \mu_2^{(B)} x^2 + \ldots$, respectively. The Lie operator is defined by $D = D_1 + hD_2 + h^2 D_3 + \ldots$, where $(D_j\Psi)(y,v) = \Psi_y(y,v) f_j(y,v) + \Psi_z(y,v) g_j(y,v)$, and it corresponds to the time derivative of solutions of (3.7).

Multistep Methods for Second Order Differential Equations. The method (1.8) for differential equations $\ddot y = f(y)$ can be treated in a similar way. In the absence of derivative approximations we get a modified differential equation of the second order

$$\ddot{y} = f(y) + hf_2(y,\dot{y}) + \ldots + h^{N-1}f_N(y,\dot{y}), \tag{3.9}$$

where the perturbation terms also depend on $\dot{y}$. Its exact solution satisfies the multistep relation (1.8) up to a defect of size $\mathcal{O}(h^{N+2})$, if (3.9) is equivalent to

$$\ddot{y} = (1 + \mu_1 hD + \mu_2 h^2 D^2 + \ldots)f(y) + \mathcal{O}(h^N), \tag{3.10}$$

where $x^2\sigma(e^x)/\rho(e^x) = 1 + \mu_1 x + \mu_2 x^2 + \ldots$, and the time derivative is given by the Lie operator $D = D_1 + hD_2 + h^2 D_3 + \ldots$ with $(D_1\Psi)(y,\dot{y}) = \Psi_y(y,\dot{y})\dot{y} + \Psi_{\dot{y}}(y,\dot{y})f(y)$ and $(D_j\Psi)(y,\dot{y}) = \Psi_{\dot{y}}(y,\dot{y})f_j(y,\dot{y})$ for $j \geq 2$. A comparison of equal powers of h in (3.9) and (3.10) uniquely defines the coefficient functions $f_j(y,\dot{y})$.

If the multistep method (1.8) is complemented with a difference formula for approximations of the derivative $v = \dot{y}$ at grid points,

$$v_n = \frac{1}{h}\sum_{j=-l}^{l} \delta_j y_{n+j}, \tag{3.11}$$

we get an additional modified differential equation

$$v = (1 + \nu_1 hD + \nu_2 h^2 D^2 + \ldots)\dot{y}. \tag{3.12}$$

The coefficients ν_j are given by $x^{-1}\delta(e^x) = 1 + \nu_1 x + \nu_2 x^2 + \ldots$, where $\delta(\zeta) = \sum_{j=-l}^{l} \delta_j \zeta^j$. For given y, this relation gives a formal one-to-one correspondence between v and $\dot{y}$. Consequently, the differential equation (3.10) combined with (3.12) can be considered as a first order differential system for the variables y and v.

XV.3.2 Parasitic Modified Equations

In practice, due to the necessity of starting approximations $y_1, \ldots, y_{k-1}$, the numerical solution of a multistep method does not lie on a solution of (3.1). For methods, where initial perturbations are not damped out sufficiently fast (cf. property (2.3) of asymptotic phase), an additional investigation is therefore needed for the study of the propagation of perturbations in the starting approximations. Let us start with two illustrating numerical experiments.

Example 3.3. Consider the explicit, linear 3-step method

$$y_{n+3} - y_{n+2} + y_{n+1} - y_n = h(f_{n+2} + f_{n+1}), \tag{3.13}$$

with characteristic polynomial $\rho(\zeta) = (\zeta - 1)(\zeta^2 + 1)$, and apply it to the pendulum equation (I.1.13). For a better illustration of the propagation of errors we consider starting approximations y_1, y_2 that are rather far from the exact solution passing through y_0. The result is shown in Fig. 3.1. We observe that the numerical solution does not lie on one smooth curve, but on four curves, and every fourth solution approximation is on the same curve.

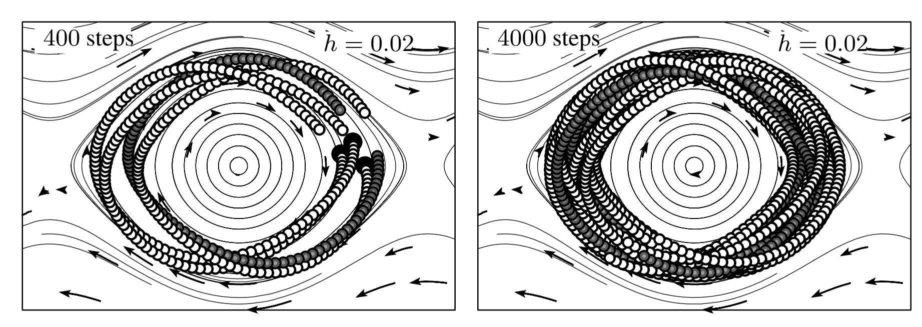


Fig. 3.1. Numerical solution of (3.13) applied to the pendulum equation. The initial approximations $y_0 = (1.9, 0.4)$, $y_1 = (1.7, 0.2)$, $y_2 = (2.1, 0)$ are indicated by black bullets; the solution points $y_3, y_7, y_{11}, \ldots$ in grey

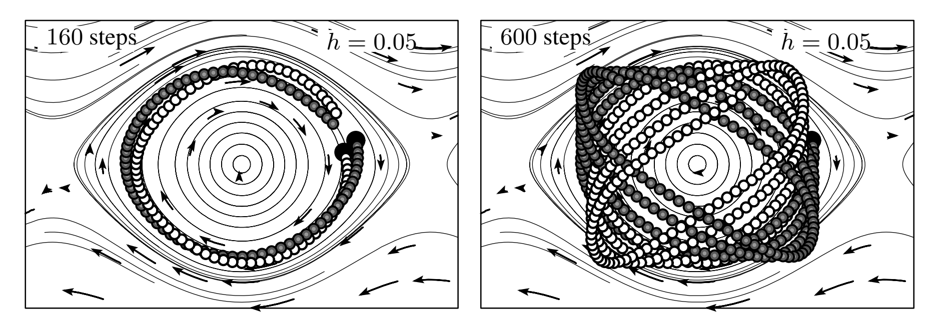


Fig. 3.2. Numerical solution of the explicit midpoint rule (3.14) applied to the pendulum equation. The initial approximations $y_0 = (1.9, 0.4)$, $y_1 = (1.7, 0.2)$ are indicated by black bullets; the solution points $y_2, y_4, y_6, \ldots$ in grey

This example shows an unexpected good long-time behaviour. Although the starting approximations are far from a smooth solution, the distance of the numerical approximations to a smooth solution curve does not increase. This is, however, not the typical situation as can be seen from our next experiment.

Example 3.4. We consider the explicit midpoint rule

$$y_{n+2} - y_n = 2h\, f_{n+1}, \tag{3.14}$$

which has $\rho(\zeta) = (\zeta - 1)(\zeta + 1)$ as characteristic polynomial. This time, the numerical solution (see Fig. 3.2) lies on two smooth curves. In contrast to the previous example, an unacceptable linear growth of the perturbations can be observed.

To be able to explain this behaviour of the multistep solutions, we complement the analysis of the modified equation for smooth numerical solutions with so-called parasitic modified equations. This theory has been developed by Hairer (1999) for first order differential equations, and extended to second order systems by Hairer & Lubich (2004).

Consider a stable, symmetric multistep method (1.1) and denote the zeros of its characteristic polynomial $\rho(\zeta)$ by $\zeta_1 = 1$ (principal root) and $\zeta_2, \ldots, \zeta_k$ (parasitic roots). We then enumerate the set of all finite products,

$$\{\zeta_\ell\}_{\ell\in\mathcal{I}} = \{\zeta = \zeta_1^{m_1} \cdot \ldots \cdot \zeta_k^{m_k}\ ;\ m_j \geq 0\} = \{\zeta_1, \ldots, \zeta_k, \zeta_{k+1}, \ldots\}. \quad (3.15)$$

It is $\{1, i, -i, -1\}$ for method (3.13) and $\{1, -1\}$ for the explicit midpoint rule (3.14). The set of subscripts $\mathcal{I}$ can be finite or infinite. We let $\mathcal{I}^* = \mathcal{I} \setminus \{1\}$, and we denote by $\mathcal{I}_N^*$ and $\mathcal{I}_N$ the finite subsets of elements which, in the representation (3.15), have $\sum_j m_j < N$.

Motivated by the previous examples and by representations of the asymptotic expansion of the global error of weakly stable multistep methods (see for example Sect. III.9 of Hairer, Nørsett & Wanner, 1993), we aim at writing the general solution y_n of the multistep method (1.1) in the form

$$y_n = y(nh) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^n z_\ell(nh), \quad (3.16)$$

where $y(t)$ and $z_\ell(t)$ are smooth functions (with derivatives bounded independently of h). The following result extends Theorem 3.1.

Theorem 3.5. *Consider a stable, consistent, and symmetric multistep method (1.1). For every truncation index $N \geq 2$, there then exist h-independent functions $f_{\ell,j}(y, \mathbf{z}^*)$ with $\mathbf{z}^* = (z_\ell)_{\ell=2}^k$ such that for every solution of*

$$\begin{aligned}
\dot y &= f_{1,1}(y,\mathbf{z}^*) + hf_{1,2}(y,\mathbf{z}^*) + \ldots + h^{N-1}f_{1,N}(y,\mathbf{z}^*) \\
\dot z_\ell &= f_{\ell,1}(y,\mathbf{z}^*) + hf_{\ell,2}(y,\mathbf{z}^*) + \ldots + h^{N-1}f_{\ell,N}(y,\mathbf{z}^*) \quad \text{for } 2 \leq \ell \leq k \\
z_\ell &= hf_{\ell,2}(y,\mathbf{z}^*) + \ldots + h^{N}f_{\ell,N+1}(y,\mathbf{z}^*) \quad \text{for } \ell > k \\
z_\ell &= 0 \quad \text{for } \ell \notin \mathcal{I}_N
\end{aligned} \quad (3.17)$$

with initial values $z_\ell(0) = \mathcal{O}(h)$ for $2 \leq \ell \leq k$, the function

$$x(t) = y(t) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^{t/h} z_\ell(t) \quad (3.18)$$

satisfies

$$\sum_{j=0}^k \alpha_j x(t+jh) = h\sum_{j=0}^k \beta_j f\big(x(t+jh)\big) + \mathcal{O}(h^{N+1}). \quad (3.19)$$

For $\mathbf{z}^* = 0$ the differential equation for y is the same as that of Theorem 3.1. The solutions of (3.17) satisfy $z_\ell(t) = \overline{z_j}(t)$ whenever $\zeta_\ell = \overline{\zeta_j}$ and this relation holds for the initial values. Moreover, $z_\ell(t) = \mathcal{O}(h^{m+1})$ on bounded time intervals if ζ_ℓ is a product of no fewer than $m \geq 2$ roots of $\rho(\zeta)$.

Proof. We let $z_1(t) := y(t)$ and insert the finite sum (3.18) into (3.19). This yields

$$\begin{aligned}\sum_{j=0}^{k}\alpha_j x(t+jh) &= \sum_{j=0}^{k}\alpha_j\sum_{\ell\in\mathcal{I}}\zeta_\ell^{(t+jh)/h}e^{jhD}z_\ell(t)\\ &= \sum_{\ell\in\mathcal{I}}\zeta_\ell^{t/h}\sum_{j=0}^{k}\alpha_j\zeta_\ell^{j}e^{jhD}z_\ell(t) \;=\; \sum_{\ell\in\mathcal{I}}\zeta_\ell^{t/h}\rho(\zeta_\ell e^{hD})z_\ell(t),\end{aligned}$$

where, as usual, D represents differentiation with respect to time. We then expand $f\big(x(t)\big)$ into a Taylor series around $y(t)$,

$$\begin{aligned}f(x(t)) &= \sum_{m\ge 0}\frac{1}{m!}f^{(m)}(y(t))\Big(\sum_{\ell_1\in\mathcal{I}^*}\zeta_{\ell_1}^{t/h}z_{\ell_1}(t),\dots,\sum_{\ell_m\in\mathcal{I}^*}\zeta_{\ell_m}^{t/h}z_{\ell_m}(t)\Big)\\ &= \sum_{\ell\in\mathcal{I}}\zeta_\ell^{t/h}\sum_{m\ge 0}\frac{1}{m!}\sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell}f^{(m)}(y(t))\big(z_{\ell_1}(t),\dots,z_{\ell_m}(t)\big)\;.\end{aligned}$$

This gives, as above,

$$\begin{aligned}&\sum_{j=0}^{k}\beta_j f\big(x(t+jh)\big) \qquad (3.20)\\ &=\sum_{\ell\in\mathcal{I}}\zeta_\ell^{t/h}\sigma(\zeta_\ell e^{hD})\sum_{m\ge 0}\frac{1}{m!}\sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell}f^{(m)}\big(y(t)\big)\big(z_{\ell_1}(t),\dots,z_{\ell_m}(t)\big).\end{aligned}$$

Comparing coefficients of $\zeta_\ell^{t/h}$ for $\ell\in\mathcal{I}_N$ in (3.19) thus yields

$$\rho(\zeta_\ell e^{hD})z_\ell = h\sigma(\zeta_\ell e^{hD})\sum_{m\ge 0}\frac{1}{m!}\sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell}f^{(m)}(y)\big(z_{\ell_1},\dots,z_{\ell_m}\big) \qquad (3.21)$$

(for $\ell=1$ and $m=0$ the sum is understood to include the term $f(y)$). With the expansion $x\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)=\mu_{\ell,0}+\mu_{\ell,1}x+\mu_{\ell,2}x^2+\dots$ for $1\le\ell\le k$, where ζ_ℓ is a simple root of $\rho(\zeta)$, this equation becomes

$$\dot z_\ell = \Big(\mu_{\ell,0}+\mu_{\ell,1}hD+\dots\Big)\sum_{m\ge 0}\frac{1}{m!}\sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell}f^{(m)}(y)\big(z_{\ell_1},\dots,z_{\ell_m}\big), \qquad (3.22)$$

and with $\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)=\mu_{\ell,0}+\mu_{\ell,1}x+\mu_{\ell,2}x^2+\dots$ for $\ell>k$, where $\rho(\zeta_\ell)\ne 0$,

$$z_\ell = h\Big(\mu_{\ell,0}+\mu_{\ell,1}hD+\dots\Big)\sum_{m\ge 0}\frac{1}{m!}\sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell}f^{(m)}(y)\big(z_{\ell_1},\dots,z_{\ell_m}\big). \qquad (3.23)$$

In the usual way (elimination of the first and higher derivatives by the differential equations and by the differentiated third relation of (3.17)) this allows us to define recursively the functions $f_{\ell,j}(y,\mathbf{z}^*)$.

From this construction process it follows that on bounded time intervals we have $z_\ell(t)=\mathcal{O}(h)$ for all $\ell\ge 2$, and $z_\ell(t)=\mathcal{O}(h^{m+1})$ if ζ_ℓ is a product of no fewer than $m\ge 2$ roots of $\rho(\zeta)$. In (3.20) and in the above construction of the coefficient functions $f_{\ell,j}(y,\mathbf{z}^*)$ we have neglected terms that contain at least N factors z_j. This gives rise to the $\mathcal{O}(h^{N+1})$ term in (3.19). □

Initial values $y(0)$, $z_\ell(0), \ell = 2,\ldots,k$, for the system (3.17) are obtained from the starting approximations $y_0,\ldots,y_{k-1}$ via the relation

$$y_j \;=\; y(jh) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^j z_\ell(jh), \quad j = 0,1,\ldots,k-1. \tag{3.24}$$

For $h = 0$ this represents a linear Vandermonde system for $y(0)$, $z_\ell(0)$. The Implicit Function Theorem thus proves the local existence of a solution of (3.24) for sufficiently small step sizes h. If $y_j, j = 2,\ldots,k$, approximate a solution $y_{ex}(t)$ of $\dot y = f(y)$ with an error $\mathcal{O}(h^s)$ (with $s \le r+1$, where r is the order of the method), then $y(0) - y_{ex}(0) = \mathcal{O}(h^s)$ and $z_\ell(0) = \mathcal{O}(h^s)$ for $\ell = 2,\ldots,k$.

The representation (3.16) of the numerical solution and the (principal and parasitic) modified equations (3.17) will be the main ingredients for the study of long-term stability of multistep methods in Sect. XV.5. An extension of the previous theorem to partitioned multistep methods is more or less straightforward. We leave the details as an exercise for the reader.

Multistep Methods for Second Order Differential Equations. A completely analogous result can be proved for stable, symmetric multistep methods (1.8) applied to $\ddot y = f(y)$. We again denote the zeros of $\rho(\zeta)$ by $\zeta_1 = 1$ and $\zeta_\ell, \ell = 2,\ldots,q$. Notice, however, that $\zeta_1 = 1$ is always a double zero, and the others can be simple or double zeros of $\rho(\zeta)$, and that $q \le k$. We consider the index sets $\mathcal{I}$, $\mathcal{I}^*$, $\mathcal{I}_N$, and $\mathcal{I}_N^*$ as in (3.15).

Theorem 3.6. *Consider a stable, consistent, and symmetric multistep method (1.8). For every truncation index $N \ge 2$, there then exist h-independent functions $f_{\ell,j}(y,\dot y,\mathbf{z}^*)$ (where $\mathbf{z}^*$ denotes the vector collecting as elements $z_\ell, \dot z_\ell$ if ζ_ℓ is a double root, and z_ℓ if ζ_ℓ is a simple root of $\rho(\zeta)$) such that for every solution of*

$$\begin{aligned}
\ddot y \;&=\; f_{1,1}(y,\dot y,\mathbf{z}^*) + hf_{1,2}(y,\dot y,\mathbf{z}^*) + \ldots + h^{N-1} f_{1,N}(y,\dot y,\mathbf{z}^*) \\
\ddot z_\ell \;&=\; f_{\ell,1}(y,\dot y,\mathbf{z}^*) + \ldots + h^{N-1} f_{\ell,N}(y,\dot y,\mathbf{z}^*) \qquad \textit{if } \rho(\zeta_\ell) = \rho'(\zeta_\ell) = 0 \\
\dot z_\ell \;&=\; hf_{\ell,2}(y,\dot y,\mathbf{z}^*) + \ldots + h^N f_{\ell,N+1}(y,\dot y,\mathbf{z}^*) \qquad \textit{if } \rho(\zeta_\ell) = 0,\ \rho'(\zeta_\ell) \ne 0 \\
z_\ell \;&=\; h^2 f_{\ell,3}(y,\dot y,\mathbf{z}^*) + \ldots + h^{N+1} f_{\ell,N+2}(y,\dot y,\mathbf{z}^*) \qquad \textit{if } \rho(\zeta_\ell) \ne 0 \\
z_\ell \;&=\; 0 \qquad \textit{for } \ell \notin \mathcal{I}_N
\end{aligned} \tag{3.25}$$

with initial values $z_\ell(0) = \mathcal{O}(h)$ for $2 \le \ell \le q$, the function

$$x(t) = y(t) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^{t/h} z_\ell(t) \tag{3.26}$$

satisfies

$$\sum_{j=0}^k \alpha_j x(t+jh) \;=\; h^2 \sum_{j=0}^k \beta_j f\big(x(t+jh)\big) \;+\; \mathcal{O}(h^{N+2}). \tag{3.27}$$

For $\mathbf{z}^* = 0$ the differential equation for y is the same as in (3.9). The solutions of (3.25) satisfy $z_\ell(t) = \overline{z_j}(t)$ whenever $\zeta_\ell = \overline{\zeta_j}$ and this relation holds for the initial values. Moreover, $z_\ell(t) = \mathcal{O}(h^{m+2})$ on bounded time intervals if ζ_ℓ is a product of no fewer than $m \ge 2$ roots of $\rho(\zeta)$.

Proof. In complete analogy to the proof of Theorem 3.5 we obtain

$$\rho(\zeta_\ell e^{hD})z_\ell = h^2\sigma(\zeta_\ell e^{hD}) \sum_{m\geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell} f^{(m)}(y)\big(z_{\ell_1},\ldots,z_{\ell_m}\big) \tag{3.28}$$

which differs from (3.21) only in the factor h^2. Depending on whether ζ_ℓ is a double, a simple, or not a zero of $\rho(\zeta)$, we expand $x^2\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ or $x\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ or $\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ into a series of powers of x, and we denote its coefficients by $\mu_{\ell,j}$. This then yields

$$\ddot z_\ell = \Big(\mu_{\ell,0} + \mu_{\ell,1}hD + \ldots\Big) \sum_{m\geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell} f^{(m)}(y)\big(z_{\ell_1},\ldots,z_{\ell_m}\big), \tag{3.29}$$

if $\rho(\zeta_\ell) = \rho'(\zeta_\ell) = 0$, but $\rho''(\zeta_\ell) \neq 0$ (in particular for $\ell = 1$ and $\zeta_1 = 1$),

$$\dot z_\ell = h\Big(\mu_{\ell,0} + \mu_{\ell,1}hD + \ldots\Big) \sum_{m\geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell} f^{(m)}(y)\big(z_{\ell_1},\ldots,z_{\ell_m}\big), \tag{3.30}$$

if $\rho(\zeta_\ell) = 0$, but $\rho'(\zeta_\ell) \neq 0$, and

$$z_\ell = h^2\Big(\mu_{\ell,0} + \mu_{\ell,1}hD + \ldots\Big) \sum_{m\geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1}\cdots\zeta_{\ell_m}=\zeta_\ell} f^{(m)}(y)\big(z_{\ell_1},\ldots,z_{\ell_m}\big), \tag{3.31}$$

if $\rho(\zeta_\ell) \neq 0$. The rest of the proof is identical to that of Theorem 3.5. □

For the system of modified equations (3.25) we need initial values $y(0), \dot y(0)$, $z_\ell(0), \dot z_\ell(0)$ if ζ_ℓ is a double root of $\rho(\zeta)$, and $z_\ell(0)$ if ζ_ℓ is a simple root. These initial values can be obtained from the starting approximations $y_0,\ldots,y_{k-1}$ via the relation (3.24).

Lemma 3.7. *Consider a stable, symmetric multistep method (1.8) of order r, and let the starting approximations $y_0,\ldots,y_{k-1}$ satisfy $y_j - y_{ex}(jh) = \mathcal{O}(h^s)$ with $2 \leq s \leq r+2$. Then there exist (locally) unique initial values for the system (3.25) such that its solution exactly satisfies (3.24).*

These initial values satisfy $z_\ell(0) = \overline{z_j}(0)$ if $\zeta_\ell = \overline{\zeta_j}$, and

$$\begin{aligned} &y(0) - y_{ex}(0) = \mathcal{O}(h^s), \quad h\dot y(0) - h\dot y_{ex}(0) = \mathcal{O}(h^s),\\ &z_\ell(0) = \mathcal{O}(h^s), \quad h\dot z_\ell(0) = \mathcal{O}(h^s) \qquad \textit{if } \zeta_\ell \textit{ is a double root,}\\ &z_\ell(0) = \mathcal{O}(h^s), \qquad \textit{if } \zeta_\ell \textit{ is a simple root.}\end{aligned} \tag{3.32}$$

Proof. We scale the derivatives by h, and consider $y(0)$, $h\dot y(0)$, $z_\ell(0)$ and $hz_\ell(0)$ as unknowns in the system (3.24), where $y(t)$ and $z_\ell(t)$ are a solution of (3.25). For $h = 0$ a linear, confluent Vandermonde system is obtained. Since this is an invertible matrix, the Implicit Function Theorem proves the statement. □

XV.4 Can Multistep Methods be Symplectic?

Readers might be astonished to find a question mark in the title. The reason is that we shall present two definitions of symplecticity of multistep methods applied to a Hamiltonian system

$$\dot{p} = -H_q(p,q), \qquad \dot{q} = H_p(p,q). \tag{4.1}$$

One works in the phase space of the exact flow, the other in a higher dimensional space. But which one is suitable? We further show that certain multistep methods can preserve energy over long times, even if they are not symplectic.

XV.4.1 Non-Symplecticity of the Underlying One-Step Method

A conjecture due to Feng Kang. (Y.-F. Tang 1993)

A natural definition of symplecticity consists of the requirement that the underlying one-step method (Theorem 2.4) be symplectic. This means that the (truncated) modified equation (3.1) is Hamiltonian. Unfortunately, we have the following negative result.

Theorem 4.1 (Tang 1993). *The underlying one-step method of a consistent linear multistep method (1.1) cannot be symplectic.*

Proof. We show that the first perturbation term in the modified equation (3.1) is in general not Hamiltonian. From (3.4) we know that $f_{r+1}(y) = \mu_r (D_1^r f)(y)$ which (omitting the non-zero error constant μ_r) is given by

$$\sum_{\tau\in T, |\tau|=r+1} \alpha(\tau)\, F(\tau)(y) \;=\; |\tau|! \sum_{\tau\in T, |\tau|=r+1} \frac{1}{\sigma(\tau)}\, b(\tau)\, F(\tau)(y) \tag{4.2}$$

with $b(\tau) = 1/\gamma(\tau)$ for $|\tau| = r+1$ (Theorem III.1.3 and (III.1.27)). Suppose now that (4.2) is Hamiltonian for all separable Hamiltonian vector fields $f(y) = J^{-1}\nabla H(y)$. Theorem IX.10.4 then implies

$$b(u\circ v) + b(v\circ u) = 0 \qquad \text{for all } u,v\in T \text{ with } |u|+|v| = r+1\,.$$

This, however, is in contradiction with

$$\frac{1}{\gamma(u\circ v)} + \frac{1}{\gamma(v\circ u)} = \frac{1}{\gamma(u)}\cdot\frac{1}{\gamma(v)},$$

which is a consequence of Theorem VI.7.6 (because the exact solution is a symplectic transformation and, as a B-series, has coefficients $a(\tau) = 1/\gamma(\tau)$). □

A similar negative result holds for a much larger class of integration methods. For example, it is proved by Hairer & Leone (1998) that, among the class of one-leg methods (see (4.7) below), only the implicit mid-point rule is symplectic (in the sense that the underlying one-step method is symplectic).

Partitioned Linear Multistep Methods. We know at least one symplectic method of the form (1.14). It is the symplectic Euler method (VI.3.1), which combines the implicit and the explicit Euler methods. However, we do not have better within the class of partitioned multistep methods as is shown in the next theorem.

Theorem 4.2. *If the underlying one-step method of a consistent, partitioned linear multistep method (1.14) is symplectic for all separable Hamiltonian systems, then its order satisfies* $r \le 1$.

Proof. Suppose that the order of the method is $r \ge 2$. By (3.8), the dominant perturbation term in the modified differential equation is $\mu_r^{(A)} h^r (D_1^r f)(y,z)$ for the y-component and $\mu_r^{(B)} h^r (D_1^r g)(y,z)$ for the z-component (at least one of the coefficients $\mu_r^{(A)}$ and $\mu_r^{(B)}$ is non-zero). This is a P-series with coefficients $b(\tau) = \mu_r^{(A)}/\gamma(\tau)$ if $\tau \in TP_p, |\tau| = r+1$ and $b(\tau) = \mu_r^{(B)}/\gamma(\tau)$ if $\tau \in TP_q, |\tau| = r+1$. If the underlying one-step method is symplectic (i.e., the modified differential equation is locally Hamiltonian), Theorem IX.10.4 implies that

$$b(u \circ v) + b(v \circ u) = 0 \quad \text{for} \quad u \in TP_p,\ v \in TP_q,\ |u| + |v| = r+1. \tag{4.3}$$

Taking for $u \in TP_p$ the tree with one vertex, and for $v \in TP_q$ an arbitrary tree with $|v| = r$, condition (4.3) gives the first relation of

$$\frac{\mu_r^{(A)}}{(r+1)\gamma(v)} + \frac{\mu_r^{(B)} r}{(r+1)\gamma(v)} = 0, \qquad \frac{\mu_r^{(B)}}{(r+1)\gamma(v)} + \frac{\mu_r^{(A)} r}{(r+1)\gamma(v)} = 0.$$

Exchanging the colour of the vertices gives the second relation. This contradicts our assumption $r \ge 2$. □

If we restrict our considerations to Hamiltonian systems with

$$H(p,q) = \frac{1}{2}\, p^T C p + c^T p + U(q), \tag{4.4}$$

where the kinetic energy is at most quadratic in p, we can find symplectic, partitioned multistep methods of order two. Indeed, the combination of the trapezoidal rule with the explicit midpoint rule

$$p_{n+1} - p_n = -\frac{h}{2}\Big(\nabla U(q_{n+1}) + \nabla U(q_n)\Big), \quad q_{n+1} - q_{n-1} = 2h\big(Cp_n + c\big) \tag{4.5}$$

has the Störmer–Verlet method as underlying one-step method. This is seen as follows: since the Hamiltonian is separable, formula (VI.3.4) yields the first formula of (4.5). The second relation is a consequence of $q_{n+1} - q_n + h(Cp_{n+1/2} + c)$ and $p_{n+1/2} + p_{n-1/2} = 2p_n$, and uses the linearity of $H_p(p,q)$.

Also for this special class of Hamiltonian systems we cannot achieve high order and symplecticity at the same time.

Theorem 4.3. *If the underlying one-step method of a consistent, partitioned linear multistep method (1.14) is symplectic for all Hamiltonian systems with Hamiltonian of the form (4.4), then its order satisfies* $r \le 2$.

Proof. The beginning is the same as that for Theorem 4.2. We let $r \geq 2$ be the order of the method (A) so that $\mu_r^{(A)} \neq 0$. Instead of (4.3) we now have to use

$$b(u \circ\circ v) - b(v \circ\circ u) = 0 \qquad \text{for} \quad u, v \in TN_p,\ |u| + |v| = r, \tag{4.6}$$

which also follows from Theorem IX.10.4. Taking for $u \in TN_p$ the tree with one vertex, and for $v \in TN_p$ an arbitrary tree with $|v| = r - 1$, condition (4.6) gives the relation

$$\frac{\mu_r^{(A)}(r-1)}{2(r+1)\gamma(v)} - \frac{\mu_r^{(A)}}{r(r+1)\gamma(v)} = 0,$$

which is contradictory for $r > 2$, because $\mu_r^{(A)} \neq 0$. □

Remark 4.4. We believe that the statement of Theorem 4.3 remains true, if we restrict our consideration to Hamiltonian functions (4.4) with $c = 0$ and invertible matrix C. Since multistep methods (1.8) for second order differential equations can be converted into partitioned multistep methods, this then implies that methods (1.8) cannot be symplectic unless the order satisfies $r \leq 2$.

XV.4.2 Symplecticity in the Higher-Dimensional Phase Space

We present here a second approach for the definition of symplecticity of multistep methods (more precisely, of one-leg methods). It is much inspired by the G-stability theory of Dahlquist (1975) for the study of stiff differential equations.

To simplify the nonlinear stability theory of linear multistep methods (1.1), Dahlquist (1975) introduced the so-called *one-leg methods*, which are defined by the relation

$$\sum_{j=0}^{k} \alpha_j y_{n+j} \;=\; hf\Big(\sum_{j=0}^{k} \beta_j y_{n+j}\Big), \tag{4.7}$$

where the normalization $\sigma(1) = \sum_j \beta_j = 1$ is assumed. In fact, there is a close relationship between the numerical solution of (4.7) and (1.1), and their long-time behaviour is the same (see Sect. V.6 of Hairer & Wanner, 1996). In the following we consider the super-vectors $Y_n = (y_{n+k-1}, \ldots, y_n)^T$ collecting k consecutive approximations of the solution.

Definition 4.5. Let G be an invertible symmetric matrix of dimension k. A k-step multistep or one-leg method is called *G-symplectic* if

$$Y_{n+1}^T\,(G \otimes S)\,Y_{n+1} = Y_n^T\,(G \otimes S)\,Y_n, \tag{4.8}$$

whenever the differential equation $\dot y = f(y)$ has $y^T S y$ as invariant (with symmetric S), i.e., the vector field satisfies $y^T S f(y) = 0$ for all y.

It is of course also possible to express this definition in terms of differential forms. As a consequence of Lemma VI.4.1 the conservation of quadratic first integrals is equivalent to symplecticity (Bochev & Scovel 1994).

In contrast to the negative results of Sect. XV.4.1, there exist a lot of G-symplectic methods. We have the following result.

Theorem 4.6 (Eirola & Sanz-Serna 1992). *Every irreducible symmetric one-leg method (4.7) is G-symplectic for some matrix G.*

Proof. We recall that a one-leg method is irreducible if the generating polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ have no common zeros.

Construction of G. The symmetry relation (1.3) implies $\rho(1/\zeta) = -\zeta^{-k}\rho(\zeta)$ and $\sigma(1/\zeta) = \zeta^{-k}\sigma(\zeta)$. Consequently, the polynomial $\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)$ vanishes for $\omega = 1/\zeta$, and contains the factor $\zeta\omega - 1$. We then define G by

$$\frac{1}{2}\big(\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)\big) \;=\; (\zeta\omega - 1)\sum_{i,j=1}^{k} g_{ij}\zeta^{i-1}\omega^{j-1}. \tag{4.9}$$

The matrix G obtained in this way is symmetric.

Regularity of G. Applying the geometric series we get

$$\sum_{i,j=1}^{k} g_{ij}\zeta^{i-1}\omega^{j-1} \;=\; -\frac{1}{2}\big(\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)\big)\big(1 + \zeta\omega + \zeta^2\omega^2 + \ldots\big),$$

where the identity holds as formal power series. Suppose that the matrix G is not invertible. Then there exists a vector $u = (u_0, u_1, \ldots, u_{k-1})^T$ such that $Gu = 0$. We formally replace the appearances of ω^{j-1} with u_{j-1} for $j \le k$ and with zero for $j > k$. This gives an identity of the form $0 = \rho(\zeta)a(\zeta) + \sigma(\zeta)b(\zeta)$ with polynomials $a(\zeta)$ and $b(\zeta)$ of degree at most $k-1$, and we get a contradiction with the irreducibility of the method.

G-Symplecticity. We next replace in (4.9) $\zeta^i\omega^j$ with $y_{n+i}^T S y_{n+j}$. Together with (4.7) this yields

$$h\Big(\sum_{i=0}^{k}\beta_i y_{n+i}\Big)^T S f\Big(\sum_{i=0}^{k}\beta_i y_{n+i}\Big) = Y_{n+1}^T(G\otimes S)Y_{n+1} - Y_n^T(G\otimes S)Y_n,$$

where $Y_n = (y_n, \ldots, y_{n+k-1})^T$. This proves (4.8) for all functions $f(y)$ satisfying $y^T S f(y) = 0$. □

Example 4.7. We consider the explicit midpoint rule (1.6), which is also a one-leg method, and the 3-step method (3.13). By Theorem 4.6 the one-leg versions are G-symplectic. Following the constructive proof of this theorem we find

$$G = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad G = \begin{pmatrix} 0 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

respectively. We apply both methods to two closely related Hamiltonian systems, namely the pendulum equation with $H(p,q) = p^2/2 - \cos q$ and a perturbed problem with $H(p,q) = p^2/2 - \cos q(1 - p/6)$, and we study the preservation of the

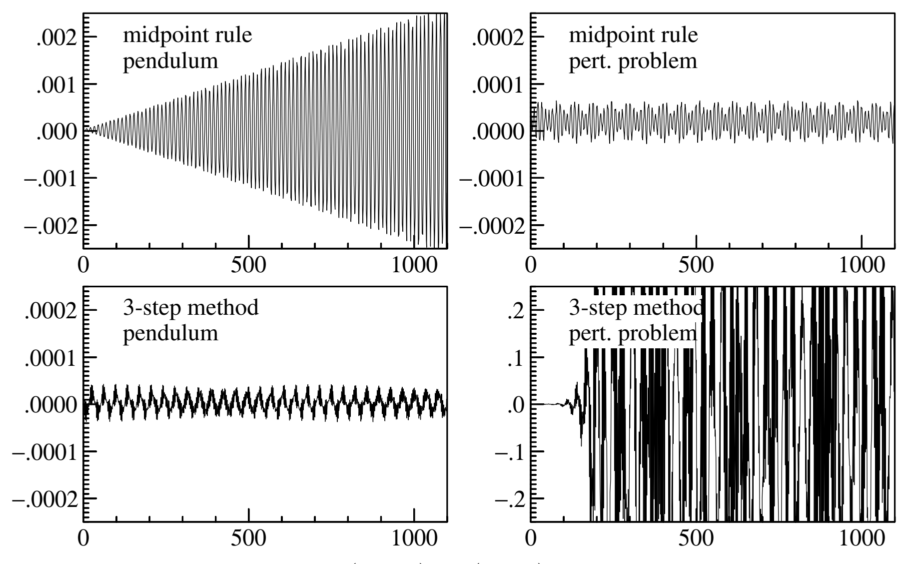


Fig. 4.1. Numerical Hamiltonian $H(p_n, q_n) - H(p_0, q_0)$ of the explicit mid-point rule and the 3-step method (3.13), applied with step size $h = 0.01$ to the pendulum problem ($H(p,q) = p^2/2 - \cos q$) and to a perturbed problem ($H(p,q) = p^2/2 - \cos q(1 - p/6)$) on the interval $[0, 1100]$ (only every 131st step is drawn)

Hamiltonian (see Fig. 4.1). The result is somewhat surprising. The midpoint rule behaves well for the perturbed problem, but shows a linear error growth in the Hamiltonian for the pendulum problem. On the other side, the weakly stable 3-step method behaves well for the pendulum equation (which is in agreement with the stable behaviour of Fig. 3.1), but has an exponential error growth for the perturbed problem. Notice that different scales are used in the four pictures.

The above example illustrates that G-symplecticity of a numerical method is not sufficient for a good long-time behaviour. It is necessary to get under control the parasitic solution components.

XV.4.3 Modified Hamiltonian of Multistep Methods

After the negative results of Sect. XV.4.1, we are fortunately also able to prove positive results concerning the near conservation of the Hamiltonian.

Theorem 4.8. *For a symmetric, consistent linear multistep method (1.1) of order r applied to $\dot y = J^{-1}\nabla H(y)$, there exists a series of the form*

$$\widetilde{H}(y) = H(y) + h^r H_{r+1}(y) + h^{r+2} H_{r+3}(y) + \ldots\ , \tag{4.10}$$

which is a formal first integral of the modified equation (3.1) without truncation.

Proof. With $\rho(e^x)/(x\sigma(e^x)) = 1 + \gamma_r x^r + \gamma_{r+2}x^{r+2} + \ldots$ it follows from (3.3) that the solution of the modified differential equation satisfies

$$(1 + \gamma_r h^r D^r + \gamma_{r+2}h^{r+2}D^{r+2} + \ldots)\dot{y} \,=\, J^{-1}\nabla H(y) + \mathcal{O}(h^N), \qquad (4.11)$$

where, due to the symmetry of the method, only odd derivatives of $y(t)$ appear. We multiply both sides with $\dot{y}^T J$ so that the right-hand side becomes the total derivative $\frac{d}{dt}H(y)$. On the left-hand side we note $\dot{y}^T J\dot{y} \,=\, 0$, $\dot{y}^T Jy^{(3)} \,=\, \frac{d}{dt}\big(\dot{y}^T J\ddot{y}\big)$ and similarly for higher derivatives

$$\dot{y}^T Jy^{(2m+1)} = \frac{d}{dt}\Big(\dot{y}^T Jy^{(2m)} - \ddot{y}^T Jy^{(2m-1)} + \ldots \pm y^{(m)T} Jy^{(m+1)}\Big). \quad (4.12)$$

We thus obtain a time derivative of an expression in which the appearing derivatives can be substituted as functions of y via the modified differential equation (3.1). Altogether this yields

$$-\frac{d}{dt}\Big(h^r H_{r+1}(y) + h^{r+2}H_{r+3}(y) + \ldots\Big) = \frac{d}{dt}H(y) + \mathcal{O}(h^N).$$

which proves the statement. □

The statement of the previous theorem is somewhat surprising. The underlying one-step method, although not symplectic, nearly conserves the Hamiltonian for general $H(y)$ (not even reversibility is required). This indicates that the condition (IX.9.20) can be satisfied for all trees also by non-symplectic methods.

For partitioned multistep methods we do not know of a similar result unless if we restrict our consideration to Hamiltonians of the form (4.4). In this case we are concerned with multistep methods for second order differential equations.

Theorem 4.9. *For a symmetric, consistent linear multistep method (1.8) of order r applied to $\ddot{y} = -\nabla U(y)$, there exists a series of the form*

$$\widetilde{H}(y,\dot{y}) = \frac{1}{2}\,\dot{y}^T\dot{y} + U(y) + h^r H_{r+1}(y,\dot{y}) + h^{r+2}H_{r+3}(y,\dot{y}) + \ldots\ , \qquad (4.13)$$

which is a formal first integral of the modified equation (3.9) without truncation.

Proof. The proof is very similar to that of the previous theorem. We expand $\rho(e^x)/(x^2\sigma(e^x)) = 1 + \gamma_r x^r + \gamma_{r+2}x^{r+2} + \ldots$, and similar to (3.10) we obtain

$$(1 + \gamma_r h^r D^r + \gamma_{r+2}h^{r+2}D^{r+2} + \ldots)\ddot{y} \,=\, -\nabla U(y) + \mathcal{O}(h^N). \qquad (4.14)$$

This time we multiply both sides with $\dot{y}^T$. The right-hand side becomes the total derivative $\frac{d}{dt}U(y)$, and for the left-hand side we use $\dot{y}^T\ddot{y} = \frac{d}{dt}(\dot{y}^T\dot{y})$ and for higher even-order derivatives

$$\dot{y}^T y^{(2m)} = \frac{d}{dt}\Big(\dot{y}^T y^{(2m-1)} - \ddot{y}^T y^{(2m-2)} + \ldots \pm \frac{1}{2}\,y^{(m)T}y^{(m)}\Big). \qquad (4.15)$$

Integrating and substituting second and higher derivatives of y via the modified differential equation (3.9) yields the desired formal first integral close to the Hamiltonian of the system. □

The formal first integral (4.13) does not depend on how approximations to the derivative $v = \dot{y}$ are obtained. If the derivative at grid points is numerically computed with the formula (3.11), then one can use the one-to-one correspondence (3.12) to express the coefficient functions of the modified differential equation in terms of y and v.

XV.4.4 Modified Quadratic First Integrals

Symplectic one-step methods exactly preserve quadratic first integrals (Sect. IV.2). This is not true for the underlying one-step method of symmetric multistep methods. However, as we shall prove in this section, it nearly preserves such first integrals.

Theorem 4.10. *Let $Q(y) = y^T C y$ (with a symmetric matrix C) be a first integral of $\dot{y} = f(y)$. For a symmetric, consistent linear multistep method (1.1) of order r, there then exists a series of the form*

$$\widetilde{Q}(y) = y^T C y + h^r Q_{r+1}(y) + h^{r+2} Q_{r+3}(y) + \ldots , \tag{4.16}$$

which is a formal first integral of the modified equation (3.1) without truncation.

Proof. We multiply (4.11) with $y^T C$ und thus obtain

$$y^T C(1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \ldots)\dot{y} \;=\; y^T C f(y) + \mathcal{O}(h^N).$$

Since $y^T C y$ is a first integral, the term on the right-hand side vanishes. For the terms on the left-hand side we notice that $y^T C \dot{y} = \frac{1}{2}\frac{d}{dt}(y^T C y)$ and that

$$y^T C y^{(2m+1)} = \frac{d}{dt}\Big(y^T C y^{(2m)} - \dot{y}^T C y^{(2m-1)} + \ldots \pm \frac{1}{2}\, y^{(m)T} C y^{(m)} \Big). \tag{4.17}$$

As in the proofs of Sect. XV.4.3 we now deduce the statement. □

A similar result holds for second order differential equations and methods (1.8). This concerns for example the total angular momentum in N-body systems.

Theorem 4.11. *Suppose that $\ddot{y} = f(y)$ has $L(y, \dot{y}) = y^T E \dot{y}$ as first integral, i.e., E is skew-symmetric and $y^T E f(y) = 0$. For a symmetric, consistent linear multistep method (1.8) of order r, there then exists a series of the form*

$$\widetilde{L}(y, \dot{y}) = y^T E \dot{y} + h^r L_{r+1}(y, \dot{y}) + h^{r+2} L_{r+3}(y, \dot{y}) + \ldots , \tag{4.18}$$

which is a formal first integral of the modified equation (3.9) without truncation.

Proof. Multiplying (4.14) with $y^T E$ gives

$$y^T E(1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \ldots)\ddot{y} \;=\; y^T E f(y) + \mathcal{O}(h^N).$$

The term at the right vanishes. Since E is a skew-symmetric matrix, we have for the terms to the left that $y^T E \ddot{y} = \frac{d}{dt} y^T E \dot{y}$ and that

$$y^T E y^{(2m+2)} = \frac{d}{dt}\Big(y^T E y^{(2m+1)} - \dot{y}^T D y^{(2m)} + \ldots \pm y^{(m)T} E y^{(m+1)} \Big). \tag{4.19}$$

This yields the statement as in the previous proofs. □

Remark 4.12. Noticing that the underlying one-step method of a symmetric multistep method can be expressed as a formal B-series (cf. Sect. XV.2.2), it follows from (4.17) that the modified first integral of Theorem 4.10 is of the form (VI.8.6). By Theorem VI.8.5 the underlying one-step method is therefore conjugate to a symplectic integrator.

A similar result holds for symmetric methods (1.8) complemented with a symmetric derivative approximation (3.11). The variables v and $\dot y$ are related via (3.12) having an expansion in even powers of h. Substituting $\dot y = \dot y(y,v)$ of this relation into the modified first integral (4.18), we obtain an expression of the form (VI.8.11). Here, the elementary differentials correspond to the system $\dot y = v$, $\dot v = f(y)$ (v has to be identified with z). Theorem VI.8.8 combined with Theorem 4.11 proves that the underlying one-step method is conjugate to a symplectic integrator.

XV.5 Long-Term Stability

The results of Sects. XV.4.3 and XV.4.4 imply the near conservation of the total energy and of the angular momentum in N-body problems for numerical solutions of the underlying one-step method of multistep methods. This, however, is of no value as long as the parasitic solutions of the multistep method are not under control. The present section is devoted to the study of the stability of numerical solutions over long time intervals.

XV.5.1 Role of Growth Parameters

The analysis of this section is based on the representation

$$y_n = y(nh) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^n\, z_\ell(nh) \tag{5.1}$$

of the numerical solution of a multistep method (cf. formula (3.16)).

Linear Multistep Methods for First Order Equations. By Theorem 3.5 the parasitic components z_ℓ (for $2 \le \ell \le k$) are the solution of a differential equation which, by (3.22), is of the form

$$\dot z_\ell = \mu_\ell f'(y) z_\ell + \ldots\ . \tag{5.2}$$

This is just the variational equation of $\dot y = f(y)$ scaled by

$$\mu_\ell = \frac{\sigma(\zeta_\ell)}{\zeta_\ell \rho'(\zeta_\ell)}, \tag{5.3}$$

which is the so-called *growth parameter* as introduced by Dahlquist (1959) and motivated there by a linear stability analysis (see Exercise 5).

We shall illustrate at the examples of Sect. XV.3.2 that the study of the truncated equation (5.2) gives already a lot of insight into the long-time behaviour of multistep methods.

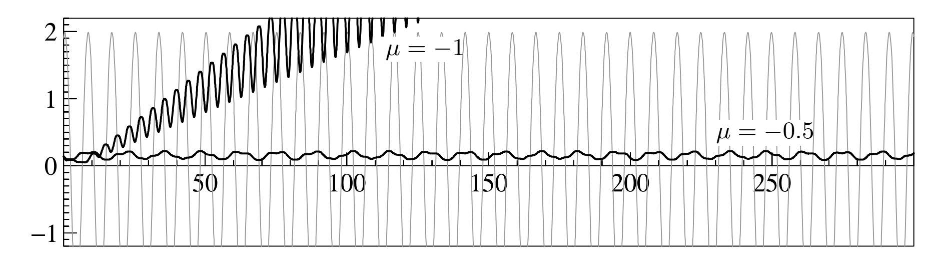


Fig. 5.1. First component of the solution of the pendulum equation (grey) together with the Euclidean norm of the solution $v(t)$ of the scaled variational equation (5.4)

Example 5.1. For the pendulum equation, the truncated equation (5.2) is

$$\begin{pmatrix} \dot y_1 \\ \dot y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\sin y_1 \end{pmatrix}, \qquad \begin{pmatrix} \dot v_1 \\ \dot v_2 \end{pmatrix} = \mu \begin{pmatrix} 0 & 1 \\ -\cos y_1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{5.4}$$

We fix initial values as $y(0) = (1.9, 0.4)^T$ and $v(0) = (0.1, 0.1)^T$. Figure 5.1 shows the solution component $y_1(t)$ in grey, and the Euclidean norm of $v(t)$ as solid black line, once for $\mu = -1$ and once for $\mu = 0.5$. We notice that the function $v(t)$ remains small and bounded for $\mu = -0.5$, and that it increases linearly for $\mu = -1$.

This agrees perfectly with the observations of Figs. 3.1 and 3.2, because the method (3.13) has growth parameter $\mu_\ell = -0.5$ for the roots $\zeta_\ell = \pm i$, whereas the explicit midpoint rule (3.14) has $\mu_\ell = -1$ for $\zeta_\ell = -1$.

The same analysis for partitioned multistep methods allows one to better understand the behaviour of the different methods in Fig. 1.3. The leading term of the parasitic modified equations depends on whether ζ_ℓ is a root of both polynomials $\rho_A(\zeta)$ and $\rho_B(\zeta)$, or only of one of them. This is very similar to the situation encountered with multistep methods for second order differential equations which we treat next.

Linear Multistep Methods for Second Order Equations. Theorem 3.6 tells us that the modified equation for the parasitic components z_ℓ depends on the multiplicity of the root ζ_ℓ. Consider a stable, symmetric method (1.8) for $\ddot y = f(y)$. If ζ_ℓ is a double root of $\rho(\zeta)$, formula (3.29) yields

$$\ddot z_\ell = \mu_\ell f'(y) z_\ell + \ldots, \qquad \mu_\ell = \frac{2\,\sigma(\zeta_\ell)}{\zeta_\ell^2\, \rho''(\zeta_\ell)}, \tag{5.5}$$

where we have not written terms containing at least two factors z_j. If ζ_ℓ is a single root of $\rho(\zeta)$, we get from (3.30) that

$$\dot z_\ell = h\mu_\ell f'(y) z_\ell + \ldots, \qquad \mu_\ell = \frac{\sigma(\zeta_\ell)}{\zeta_\ell\, \rho'(\zeta_\ell)}. \tag{5.6}$$

There is an enormous difference between the parasitic modified equations corresponding to double or single roots of $\rho(\zeta)$. Equation (5.5) is the complete analogue

of (5.2) and, as before, the long-time behaviour is hardly predictable and strongly depends on the growth parameter. For single roots, however, we are concerned with a first order differential equation (5.6) having an additional factor h as bonus. For the analysis of Sects. XV.5.2 and XV.5.3 it is important to have only single roots.

Definition 5.2. A symmetric multistep method (1.8) for second order differential equations is called *s-stable* if, apart from the double root at 1, all zeros of $\rho(\zeta)$ are simple and of modulus one.

The linearized parasitic modified equations give much insight into the long-time behaviour of multistep methods. To get rigorous estimes over long times, however, further considerations are necessary. A partial result is given by Cano & Sanz-Serna (1998) for multistep methods (1.8) applied to equations $\ddot{y} = f(y)$ with periodic exact solution. There, the first terms of the asymptotic error expansion for the global error are computed, and their growth as a function of time is studied. We shall follow the approach of Hairer & Lubich (2004) who exploit the Hamiltonian structure of second order differential equations.

XV.5.2 Hamiltonian of the Full Modified System

In the remainder of this section we restrict our consideration to s-stable, irreducible linear multistep methods

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2 \sum_{j=0}^{k} \beta_j \nabla U(q_{n+j}), \tag{5.7}$$

applied to Hamiltonian systems written as

$$\ddot{q} = -\nabla U(q), \tag{5.8}$$

where $U(q)$ is assumed to be real-analytic in the considered region.

The key to proving long-time error estimates is the observation that much of the Hamiltonian structure is conserved in the modified equations (3.25). The results and techniques of this subsection are closely related to those of Sect. XIII.6.3 developed for numerical methods for oscillatory differential equations.

We let $\mathbf{z} = (z_\ell)_{\ell \in \mathcal{I}_N}$ and define $\mathcal{U}(\mathbf{z})$ as

$$\mathcal{U}(\mathbf{z}) = U(z_0) + \sum_{m \geq 1} \frac{1}{m!} \sum_{\zeta_{\ell_1} \cdots \zeta_{\ell_m} = 1} U^{(m)}(z_0)(z_{\ell_1}, \ldots, z_{\ell_m}), \tag{5.9}$$

where the second sum is over all indices $\ell_1 \in \mathcal{I}_N^*, \ldots, \ell_m \in \mathcal{I}_N^*$ with $\zeta_{\ell_1} \ldots \zeta_{\ell_m} = 1$ (using the notation of Sect. XV.3.2). Since the roots of $\rho(\zeta)$, different from $\zeta_1 = 1$ are complex and appear in pairs (Exercise 3), also the functions z_ℓ appear in pairs. It is convenient to use the notation $z_{-\ell} = z_j$ if $\overline{\zeta_\ell} = \zeta_j$.

It follows from (3.28) with $f(q) = -\nabla U(q)$ that every solution of the truncated modified equation (3.25) satisfies

$$\rho(\zeta_\ell e^{hD}) z_\ell = -h^2 \sigma(\zeta_\ell e^{hD}) \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) + \mathcal{O}(h^{N+2}) \tag{5.10}$$

(for all $\ell \in \mathcal{I}$) as long as

$$y \in K, \quad \|\dot{y}\| \le M, \quad \|z_\ell\| \le \delta \ \text{ for } \ 1 < \ell < k, \tag{5.11}$$

where K is a compact subset of the domain of analyticity of $U(q)$, $M > 0$ some bound on the derivative, and $0 < \delta = \mathcal{O}(h)$ is a sufficiently small constant (note that this implies $\|z_\ell\| \le \delta$ for all $\ell \in \mathcal{I}^*$ if h is sufficiently small, cf. the algebraic relations of (3.25)).

For ease of presentation, we assume for the moment that $\sigma(\zeta_\ell) \ne 0$ for all $\ell \in \mathcal{I}_N$ (we know that this holds for $1 \le \ell < k$, because the method is irreducible). We apply the operator $\sigma^{-1}(\zeta_\ell e^{hD})$ to both sides of (5.10) and divide by h^2:

$$h^{-2} \Big(\frac{\rho}{\sigma}\Big)(\zeta_\ell e^{hD}) z_\ell = -\nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) + \mathcal{O}(h^N). \tag{5.12}$$

We then multiply with $\dot{z}_{-\ell}^T$, sum over all $\ell \in \mathcal{I}_N$, and thus obtain

$$h^{-2} \sum_{\ell \in \mathcal{I}_N} \dot{z}_{-\ell}^T \Big(\frac{\rho}{\sigma}\Big)(\zeta_\ell e^{hD}) z_\ell + \frac{d}{dt} \mathcal{U}(\mathbf{z}) = \mathcal{O}(h^N). \tag{5.13}$$

We now show that also the first expression on the left-hand side is a total derivative of a function depending on $\mathbf{z}$ and its time derivatives. For this we note that

$$\Big(\frac{\rho}{\sigma}\Big)(\zeta_\ell e^{ix}) = \sum_{j \ge 0} c_{\ell,j} x^j \qquad \text{with } \textit{real} \text{ coefficients} \quad c_{\ell,j} = (-1)^j c_{-\ell,j}. \tag{5.14}$$

This holds because the symmetry of the multistep method yields $(\rho/\sigma)(1/\zeta) = \overline{(\rho/\sigma)(\zeta)}$ and hence, for real x, $(\rho/\sigma)(\zeta_\ell e^{ix}) = (\rho/\sigma)(\overline{\zeta_\ell e^{ix}}) = \overline{(\rho/\sigma)(\zeta_\ell e^{ix})}$. With the expansion (5.14) we obtain

$$\Big(\frac{\rho}{\sigma}\Big)(\zeta_\ell e^{hD}) z_\ell = \sum_{j=0}^{N+1} c_{\ell,j} (-ih)^j z_\ell^{(j)} + \mathcal{O}(h^{N+2}). \tag{5.15}$$

To study (5.13) we apply the relation (4.12) for the real function $y = z_1$ and for z_ℓ corresponding to $\zeta_\ell = -1$, while for the complex-valued functions $z = z_\ell$, with complex conjugate $\overline{z} = z_{-\ell}$, we use

$$\begin{aligned} \mathrm{Re}\, \dot{\overline{z}}^T z^{(2m)} &= \mathrm{Re}\, \frac{d}{dt} \Big(\dot{\overline{z}}^T z^{(2m-1)} - \ddot{\overline{z}}^T z^{(2m-2)} + \ldots \pm \frac{1}{2} (\overline{z}^{(m)})^T z^{(m)} \Big) \\ \mathrm{Im}\, \dot{\overline{z}}^T z^{(2m+1)} &= \mathrm{Im}\, \frac{d}{dt} \Big(\dot{\overline{z}}^T z^{(2m)} - \ddot{\overline{z}}^T z^{(2m-1)} + \ldots \mp (\overline{z}^{(m)})^T z^{(m+1)} \Big). \end{aligned}$$

Together with (5.15) these relations show that the terms

$$\begin{aligned} &\dot{z}_{-\ell}^T \Big(\frac{\rho}{\sigma}\Big)(\zeta_\ell e^{hD}) z_\ell + \dot{z}_\ell^T \Big(\frac{\rho}{\sigma}\Big)(\zeta_{-\ell} e^{hD}) z_{-\ell} \\ &\qquad = \sum_{j=0}^{N+1} c_{\ell,j}\, 2\, \mathrm{Re}\Big((-ih)^j\, \dot{\overline{z}}_\ell^{\,T} z_\ell^{(j)} \Big) + \mathcal{O}(h^{N+2}) \end{aligned}$$

give a total derivative (up to the remainder term). Hence the left-hand side of (5.13) can be written as the time derivative of a function which depends on z_ℓ, $\ell \in \mathcal{I}_N$, and on their derivatives. Using the modified equation (3.25) we eliminate all z_ℓ corresponding to ζ_ℓ with $\rho(\zeta_\ell) \neq 0$ and their derivatives, the first and higher derivatives of z_ℓ (for $1 < \ell < k$), and the second and higher derivatives of $y = z_1$. We thus get a function

$$\mathcal{H}(y, \dot{y}, \mathbf{z}^*) = H_0(y, \dot{y}, \mathbf{z}^*) + \ldots + h^{N-1} H_{N-1}(y, \dot{y}, \mathbf{z}^*) \tag{5.16}$$

with $\mathbf{z}^* = (z_\ell)_{\ell=2}^{k-1}$, such that

$$\frac{d}{dt}\mathcal{H}\big(y(t), \dot{y}(t), \mathbf{z}^*(t)\big) = \mathcal{O}(h^N), \tag{5.17}$$

along solutions of (3.25) that stay in a set defined by (5.11). The function $\mathcal{H}$ is therefore an almost-invariant of the system (3.25).

If, however, $\sigma(\zeta)$ does have a zero ζ_ℓ, then we omit the corresponding term from the sum in (5.13). Hence the term $\dot{z}_{-\ell}^T \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z})$ is missing from $(d/dt)\mathcal{U}(\mathbf{z})$ and must therefore be compensated in the remainder term. Since ζ_ℓ is a product of no fewer than two zeros of $\rho(\zeta)$, it follows from (3.31) and from $\mu_{\ell,0} = 0$ that $z_\ell = \mathcal{O}(h^3\delta^2)$, as long as $\|z_j\| \leq \delta$ for $1 < j < k$. We further have $\nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) = \mathcal{O}(\delta^2)$, so that the remainder term in (5.17) is augmented by $\mathcal{O}(h^3\delta^4)$.

We summarize the above considerations (Hairer & Lubich 2004) as follows.

Theorem 5.3. *Every solution of the truncated modified equation (3.25) satisfies, with $\mathcal{H}$ from (5.16),*

$$\mathcal{H}\big(y(t), \dot{y}(t), \mathbf{z}^*(t)\big) = \mathcal{H}\big(y(0), \dot{y}(0), \mathbf{z}^*(0)\big) + \mathcal{O}(th^N) + \mathcal{O}(th^3\delta^4) \tag{5.18}$$

as long as the solution stays in the set defined by (5.11). Moreover,

$$\mathcal{H}\big(y, \dot{y}, \mathbf{z}^*\big) = H(y, \dot{y}) + \mathcal{O}(h^p) + \mathcal{O}(h\delta^2). \tag{5.19}$$

The closeness to the Hamiltonian $H(y, \dot{y}) = \frac{1}{2}\|\dot{y}\|^2 + U(y)$ follows also directly from the above construction. For $\mathbf{z}^* = 0$ we have $\mathcal{H}(y, \dot{y}, 0) = \widetilde{H}(y, \dot{y})$, where $\widetilde{H}$ is the modified energy from Theorem 4.9.

We will use Theorem 5.3 in Section XV.6.1 to infer the long-time near-conservation of the Hamiltonian along numerical solutions. Before that we need to bound the parasitic components.

XV.5.3 Long-Time Bounds for Parasitic Solution Components

The modified equations have further almost-invariants which are close to the squares of the norms of the parasitic components that correspond to the roots of $\rho(\zeta)$. We derive them here and use them to show that all parasitic solution components remain small over very long times. The techniques used in this subsection are similar to those in Sects. XIII.6 and XIII.7.

We consider ℓ with $1 < \ell < k$ for which ζ_ℓ is a *simple* root of $\rho(\zeta)$ and $\sigma(\zeta_\ell) \neq 0$. The dominant term on the left-hand side of (5.12) is $-c_{\ell,1} i h^{-1} \dot{z}_\ell$. Since

$$\frac{d}{dt}\|z_\ell\|^2 = z_{-\ell}^T \dot{z}_\ell + z_\ell^T \dot{z}_{-\ell}, \tag{5.20}$$

we multiply (5.12) with $z_{-\ell}^T$ and the corresponding equation for $\zeta_{-\ell}$ with z_ℓ^T, and we form the difference, so that the dominant term on the left-hand side becomes $-c_{\ell,1} i h^{-1} \frac{d}{dt}\|z_\ell\|^2$ (note $c_{-\ell,1} = -c_{\ell,1}$). Dividing by $-c_{\ell,1} i h^{-1}$ gives

$$\begin{aligned}\frac{i}{c_{\ell,1} h}\Big(z_{-\ell}^T \frac{\rho}{\sigma}(\zeta_\ell e^{hD}) z_\ell - z_\ell^T \frac{\rho}{\sigma}(\zeta_{-\ell} e^{hD}) z_{-\ell}\Big)& \\ = \frac{ih}{c_{\ell,1}}\Big(-z_{-\ell}^T \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) + z_\ell^T \nabla_{z_\ell} \mathcal{U}(\mathbf{z})\Big)&.\end{aligned} \tag{5.21}$$

We first estimate the right-hand expression. Since

$$\nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) = \nabla^2 U(z_0) z_\ell + \mathcal{O}(\delta^2),$$

as long as (5.11) is satisfied, we obtain from the symmetry of the Hessian that the right-hand side of (5.21) is of size $\mathcal{O}(h\delta^3)$. The dominant $\mathcal{O}(h\delta^3)$ term is present only if $\zeta_{-\ell}$ can be written as the product of two roots of $\rho(\zeta)$ other than 1. If this is not the case, the expression (5.21) is of size $\mathcal{O}(h\delta^4)$.

Using the expansion (5.15) on the left-hand side of (5.21) and the relations (for $z = z_\ell$)

$$\begin{aligned}\operatorname{Re} \overline{z}^T z^{(2m+1)} &= \operatorname{Re} \frac{d}{dt}\Big(\overline{z}^T z^{(2m)} - \dot{\overline{z}}^T z^{(2m-1)} \ldots \mp \frac{1}{2} \big(\overline{z}^{(m)}\big)^T z^{(m)}\Big) \\ \operatorname{Im} \overline{z}^T z^{(2m+2)} &= \operatorname{Im} \frac{d}{dt}\Big(\overline{z}^T z^{(2m+1)} - \dot{\overline{z}}^T z^{(2m)} + \ldots \pm \big(\overline{z}^{(m)}\big)^T z^{(m+1)}\Big)\end{aligned}$$

we obtain that (5.21) is, up to $\mathcal{O}(h^N)$, the total derivative of a function depending on $\mathbf{z}$ and its derivatives.

By construction the dominant term is $\frac{d}{dt}\|z_\ell\|^2$. The following terms have at least one more power of h and at least one derivative which by (3.25) gives rise to an additional factor h. Eliminating higher derivatives with the help of (3.25), we arrive at a function of the form

$$\mathcal{K}_\ell(y, \dot{y}, \mathbf{z}^*) = \|z_\ell\|^2 + h^2 K_{\ell,2}(y, \dot{y}, \mathbf{z}^*) + \ldots + h^{N-1} K_{\ell,N-1}(y, \dot{y}, \mathbf{z}^*). \tag{5.22}$$

As we have seen, its total derivative is of size $\mathcal{O}(h\delta^3)$ or smaller. We summarize these considerations in the following theorem.

Theorem 5.4. *Along every solution of the truncated modified equation (3.25) the function* $\mathcal{K}_\ell(y, \dot{y}, \mathbf{z}^*)$ *satisfies for* $1 < \ell < k$

$$\mathcal{K}_\ell\big(y(t), \dot{y}(t), \mathbf{z}^*(t)\big) = \mathcal{K}_\ell\big(y(0), \dot{y}(0), \mathbf{z}^*(0)\big) + \mathcal{O}(th^N) + \mathcal{O}(th\delta^3) \tag{5.23}$$

as long as the solution stays in the set defined by (5.11). The second error term is replaced by $\mathcal{O}(th\delta^4)$ if no root of $\rho(\zeta)$ other than 1 is the product of two other roots. Moreover,

$$\mathcal{K}_\ell\big(y,\dot y,\mathbf{z}^*\big) = \|z_\ell\|^2 + \mathcal{O}(h^2\delta^2). \tag{5.24}$$

This result allows us to write the numerical solution in a form that is suitable for deriving long-time error estimates. Let us first collect the necessary assumptions:

(A1) the multistep method (5.7) is symmetric, s-stable, and of order r;
(A2) the potential function $U(q)$ of (5.8) is defined and analytic in an open neighbourhood of a compact set K;
(A3) the starting approximations $q_0,\ldots,q_{k-1}$ are such that the initial values for (3.25) obtained from Lemma 3.7 satisfy $y(0)\in K$, $\|\dot y(0)\|\le M$, and $\|z_\ell(0)\|\le\delta/2$ for $1<\ell<k$;
(A4) the numerical solution $\{q_n\}$ stays for $0\le nh\le T$ in a compact set K_0 which has a positive distance to the boundary of K.

Theorem 5.5 (Hairer & Lubich 2004). *Assume (A1)–(A4). For sufficiently small h and δ and for a fixed truncation index N (large enough such that $h^N=\mathcal{O}(\delta^4)$), there exist functions $y(t)$ and $z_\ell(t)$ on an interval of length*

$$T=\mathcal{O}((h\delta)^{-1})$$

such that

- $q_n = y(nh) + \sum_{\ell\in\mathcal{I}^*}\zeta_\ell^n z_\ell(nh)$ *for $0\le nh\le T$;*
- *on every subinterval $[jh,(j+1)h)$ the functions $y(t), z_\ell(t)$ are a solution of the system (3.25);*
- *the functions $y(t), z_\ell(t)$ have jump discontinuities of size $\mathcal{O}(h^{N+2})$ at the grid points jh;*
- $\|z_\ell(t)\|\le\delta$ *for $0\le t\le T$.*

If no root of $\rho(\zeta)$ other than 1 is the product of two other roots, all these estimates are valid on an interval of length $T=\mathcal{O}((h\delta^2)^{-1})$.

Proof. To define the functions $y(t), z_\ell(t)$ on the interval $[jh,(j+1)h)$ we consider the k consecutive numerical solution values $q_j, q_{j+1},\ldots, q_{j+k-1}$. We compute initial values for (3.25) according to Lemma 3.7, and we let $y(t), z_\ell(t)$ be a solution of (3.25) on $[jh,(j+1)h)$. Because their defect is $\mathcal{O}(h^N)$ and $\mathcal{O}(h^{N+1})$, respectively, such a construction yields jump discontinuities of size $\mathcal{O}(h^{N+2})$ at the grid points.

It follows from Theorem 5.4 that $\mathcal{K}_\ell(y(t),\dot y(t),\mathbf{z}^*(t))$ remains constant up to an error of size $\mathcal{O}(h^2\delta^3)$ on the interval $[jh,(j+1)h)$. Taking into account the jump discontinuities, we find that

$$\mathcal{K}_\ell(y(t),\dot y(t),\mathbf{z}^*(t)) \le \mathcal{K}_\ell(y(0),\dot y(0),\mathbf{z}^*(0)) + C_1 th\delta^3 + C_2 th^{N+1} \tag{5.25}$$

as long as $\|z_\ell(t)\|\le\delta$. By (5.24) this then implies

$$\|z_\ell(t)\|^2 \le \|z_\ell(0)\|^2 + C_1 th\delta^3 + C_2 th^{N+1} + C_3 h^2\delta^2. \tag{5.26}$$

The assumption $\|z_\ell(t)\| \le \delta$ is certainly satisfied as long as $C_1 th\delta \le 1/4$, $C_2 th^{N+1} \le \delta^2/4$, and $C_3 h^2 \le 1/4$, so that the right-hand side of (5.26) is bounded by δ^2. This proves not only the estimate for $\|z_\ell(t)\|$, but at the same time it guarantees recursively that the above construction of the functions $y(t), z_\ell(t)$ is feasible. □

Notice that for initial values computed by a sufficiently accurate one-step method the constant δ can be chosen as small as $\mathcal{O}(h^{r+2})$ where r is the order of the multistep method (cf. Lemma 3.7). The above estimates are therefore valid on very long time intervals.

Example 5.6. To illustrate the long-time behaviour of the parasitic terms z_ℓ we consider the pendulum equation $\ddot q = -\sin q$, and we apply the symmetric multistep methods (B) and (C) of Example 1.2. For method (C), the starting values are chosen far from a smooth solution, so that the propagation of the parasitic terms in the numerical solution can be better observed. We compute the velocity approximation by

$$v_n = \frac{h}{12}\Big(8(q_{n+1} - q_{n-1}) - (q_{n+2} - q_{n-2})\Big). \tag{5.27}$$

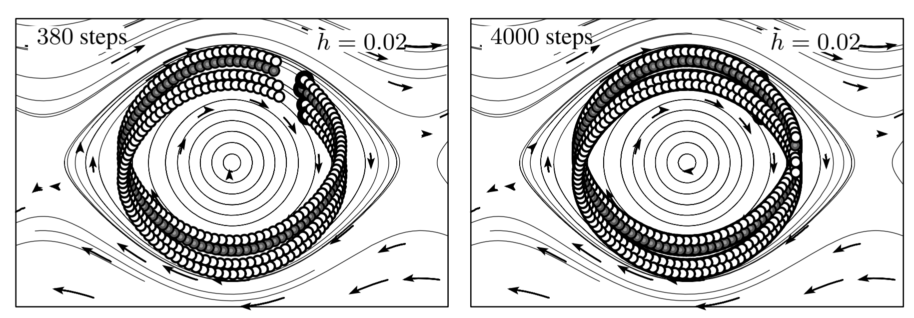


Fig. 5.2. Stable propagation of perturbations in the starting values for method (C) of Example 1.2; initial values are $q_0 = 1.141$ $q_1 = 1.158$, $q_2 = 1.178$, and $q_3 = 1.206$

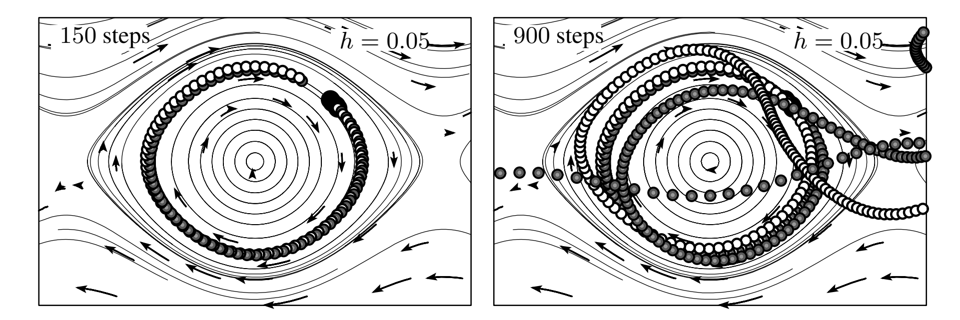


Fig. 5.3. Unstable propagation of perturbations in the starting values, for method (B) of Example 1.2; initial values are $q_0 = 1.147$ $q_1 = 1.183$, $q_2 = 1.255$, and $q_3 = 1.286$

Figure 5.2 shows the numerical solution (q_n, v_n) for $n \geq 2$. The values for $n = 2, 3, 4, 5$ are indicated by larger black bullets. The parasitic roots of method (C) are $\pm i$ and both are simple. The numerical solution is therefore of the form

$$q_n = y(nh) + i^n z_1(nh) + (-i)^n \overline{z_1(nh)} + (-1)^n z_2(nh).$$

One observes in Fig. 5.2 that the functions $z_j(t)$ not only remain bounded and small, but they stay essentially constant over the considered interval. This should be compared to Fig. 3.1, where the parasitic functions $z_j(t)$ are bounded, but not constant.

Method (B) has a double parasitic root at -1 and, therefore, is not s-stable. Its numerical solution behaves like $q_n = y(nh) + (-1)^n z(nh)$. In Fig. 5.3 every second approximation is drawn in grey. One sees that the numerical solution stays on two smooth curves $y(t) + z(t)$ and $y(t) - z(t)$ which, however, do not remain close to each other.

XV.6 Explanation of the Long-Time Behaviour

The bounds on the parasitic solution components of Sect. XV.5.3 allow us to get rigorous statements on the long-time behaviour of multistep methods (5.7) for second order differential equations. The following results are taken from Hairer & Lubich (2004). We do not know of similar results for multistep methods (1.1).

XV.6.1 Conservation of Energy and Angular Momentum

The energy conservation is now a direct consequence of Theorems 5.3 and 5.5. We shall use the representation of q_n in terms of functions $y(t), z_\ell(t)$ as in Theorem 5.5. Taking into account the jump discontinuities of these functions, Theorem 5.3 yields

$$\mathcal{H}(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{H}(y(0), \dot{y}(0), \mathbf{z}^*(0)) + \mathcal{O}(th^3\delta^4) + \mathcal{O}(th^{N+1}).$$

We have $\delta = \mathcal{O}(h^{r+1})$ if the starting approximations are computed by a rth order one-step method. If N is chosen sufficiently large, this together with (5.19) implies

$$H(y(t), \dot{y}(t)) = H(y(0), \dot{y}(0)) + \mathcal{O}(h^p) \qquad \text{for} \quad 0 \leq t \leq T = \mathcal{O}(h^{-p-2}).$$

If the velocity approximation $p_n = v_n$ is given by a rth order finite difference formula (3.11), it follows from Theorem 5.5 that $p_n = \dot{y}(nh) + \mathcal{O}(h^r)$ provided the truncation index N is sufficiently large. This proves the following result, and explains the excellent long-time behaviour of method (C) in Fig. 1.2.

Theorem 6.1 (Total Energy). *For a problem $\ddot{q} = -\nabla U(q)$ with total energy $H(p, q) = \frac{1}{2}p^T p + U(q)$, the numerical solution of an s-stable symmetric multistep method (5.7) of order r satisfies*

$$H(q_n, p_n) = H(q_0, p_0) + \mathcal{O}(h^r) \qquad \textit{for} \quad nh \leq h^{-r-2}.$$

If no root of $\rho(\zeta)$ other than 1 is a product of two other roots, the statement holds on intervals of length $\mathcal{O}(h^{-2r-3})$. □

We assume next that the differential equation $\ddot{q} = -\nabla U(q)$ has a quadratic first integral of the form $L(q,\dot{q}) = \dot{q}^T Aq$ (e.g., the angular momentum in N-body problems). This means that A is skew-symmetric and $\nabla U(q)^T Aq = 0$. The last equation can also be interpreted as the invariance relation $U(e^{\tau A}q) = U(q)$. This property implies for $\mathcal{U}(\mathbf{z})$, given by (5.9), that $\mathcal{U}(e^{\tau A}\mathbf{z}) = \mathcal{U}(\mathbf{z})$ (here $e^{\tau A}\mathbf{z} = (e^{\tau A}z_\ell)_{\ell\in\mathcal{I}}$). Along solutions $\mathbf{z}(t)$ of the modified equations (5.10) we therefore have up to terms of size $\mathcal{O}(h^N)$

$$0 = \frac{d}{d\tau}\Big|_{\tau=0} \mathcal{U}(e^{\tau A}\mathbf{z}) = \sum_{\ell\in\mathcal{I}} z_{-\ell}^T\, A\, \nabla_{z_{-\ell}}\mathcal{U}(\mathbf{z}) = \sum_{\ell\in\mathcal{I}} h^{-2} z_{-\ell}^T\, A \left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{hD}) z_\ell .$$

If $\sigma(\zeta)$ has a root ζ_ℓ, then the corresponding term is omitted from the last sum, leading to a remainder term which in the worst case is $\mathcal{O}(h^3\delta^4)$, as in Theorem 5.3. Like in the previous proofs, the last sum is, for skew-symmetric A, the total derivative of a function

$$\mathcal{L}(y,\dot{y},\mathbf{z}^*) = L_0(y,\dot{y},\mathbf{z}^*) + \ldots + h^{N-1}L_{N-1}(y,\dot{y},\mathbf{z}^*)$$

which satisfies (under the same assumptions as in Theorem 5.3)

$$\mathcal{L}\big(y(t),\dot{y}(t),\mathbf{z}^*(t)\big) = \mathcal{L}\big(y(0),\dot{y}(0),\mathbf{z}^*(0)\big) + \mathcal{O}(th^3\delta^4) + \mathcal{O}(th^{N+1})$$

and

$$\mathcal{L}\big(y,\dot{y},\mathbf{z}^*\big) = L(y,\dot{y}) + \mathcal{O}(h^p) + \mathcal{O}(\delta^2/h). \tag{6.1}$$

We therefore obtain the following result.

Theorem 6.2 (Angular Momentum). *Let $L(q,\dot{q}) = \dot{q}^T Aq$ be a first integral of $\ddot{q} = -\nabla U(q)$. The numerical solution of an s-stable symmetric multistep method (5.7) of order r then satisfies*

$$L(q_n,p_n) = L(q_0,p_0) + \mathcal{O}(h^r) \qquad \textit{for} \quad nh \le h^{-r-2}.$$

If no root of $\rho(\zeta)$ other than 1 is a product of two other roots, the statement holds on intervals of length $\mathcal{O}(h^{-2r-3})$. □

XV.6.2 Linear Error Growth for Integrable Systems

The differential equation $\ddot{q} = -\nabla U(q)$, written as $\dot{q} = v$, $\dot{v} = -\nabla U(q)$, is reversible with respect to the involution $v \mapsto -v$. Assume that it is also an integrable system in the sense of Definition XI.1.1, and denote by $a = I(q,v)$ the action variables, and by $\omega(a)$ the frequences of the system.

By Theorem 5.5, the numerical solution can be written as $q_n = y(nh) + \sum_{\ell\in\mathcal{I}^*} \zeta_\ell^n z_\ell(nh)$, where (at least locally) $y(t)$ is the solution of a modified differential equation (first equation of (3.25))

$$\ddot{y} = f_{0,0}(y,\dot{y},\mathbf{z}^*) + hf_{0,1}(y,\dot{y},\mathbf{z}^*) + \ldots + h^{N-1}f_{0,N-1}(y,\dot{y},\mathbf{z}^*) \tag{6.2}$$

which, for $\mathbf{z}^* = 0$ becomes the reversible modified differential equation (3.9). Since $z_j(t) = \mathcal{O}(\delta)$ (see Theorem 5.5) and since $\mathbf{z}^*$ appears at least quadratically in (6.2), this equation is a $\mathcal{O}(\delta^2)$ perturbation of (3.9). We are now in the position to apply the results of Lemma XI.2.1 and Theorem XI.3.1. The additional (non-reversible) perturbation of size $\mathcal{O}(\delta^2)$ in the differential equation (6.2) produces an error term of size $\mathcal{O}(t\delta^2)$ in the action variables and of size $\mathcal{O}(t^2\delta^2)$ in the angle variables. If $\delta = \mathcal{O}(h^{r+1})$, these terms are negligible with respect to those already appearing in Theorem XI.3.1. The errors due to the jump discontinuities (Theorem 5.5) are also negligible. We have thus proved the following statement.

Theorem 6.3. *Consider applying the s-stable symmetric multistep method (5.7) of order r to an integrable reversible system $\ddot q = -\nabla U(q)$ with real-analytic potential U. Suppose that $\omega^* \in \mathbb{R}^d$ satisfies the diophantine condition (X.2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \le h_0$: every numerical solution (q_n, v_n) starting with frequencies $\omega_0 = \omega(I(q_0, v_0))$ such that $\|\omega_0 - \omega^*\| \le c|\log h|^{-\nu-1}$, satisfies*

$$\begin{aligned} \|(q_n, v_n) - (q(t), v(t))\| &\le C\, t\, h^r \\ \|I(q_n, v_n) - I(q_0, v_0)\| &\le C\, h^r \end{aligned} \qquad \textit{for} \quad 0 \le t = nh \le h^{-r}.$$

The constants h_0, c, C depend on d, γ, ν and on bounds of the potential. □

XV.7 Practical Considerations

In computations with multistep methods one can observe resonance phenomena, if relatively large step sizes are used. This and the use of variable step sizes are the subject of this section.

XV.7.1 Numerical Instabilities and Resonances

> Soon after Quinlan and Tremaine's methods were published, however, Alar Toomre discovered a disturbing feature of the methods, . . .
> (G.D. Quinlan 1999)

It is a simple task to derive multistep methods of high order. Consider, for example, methods of the form (1.8) for second order differential equations $\ddot y = f(y)$. Their order is determined by the condition (1.9). We choose arbitrarily $\rho(\zeta)$ such that $\zeta = 1$ is a double zero and the stability condition is satisfied. Condition (1.9) then gives

$$\sigma(\zeta) = \rho(\zeta)/\log^2 \zeta + \mathcal{O}\big((\zeta - 1)^r\big).$$

Expanding the right-hand expression into a Taylor series at $\zeta = 1$ and truncating suitably, this yields the corresponding σ polynomial. If we take

$$\rho(\zeta) = (\zeta - 1)^2(\zeta^6 + \zeta^4 + \zeta^3 + \zeta^2 + 1), \tag{7.1}$$

Table 7.1. Symmetric multistep methods for second order problems; $k = 8$ and order $r = 8$

	SY8		SY8B		SY8C	
i	α_i	$12096\,\beta_i$	α_i	$120960\,\beta_i$	α_i	$8640\,\beta_i$
0	1	0	1	0	1	0
1	-2	17671	0	192481	-1	13207
2	2	-23622	0	6582	0	-8934
3	-1	61449	$-1/2$	816783	0	42873
4	0	-50516	-1	-156812	0	-33812

we get in this way Method SY8 of Table 7.1, a method proposed by Quinlan & Tremaine (1990) for computations in celestial mechanics. All methods of Table 7.1 are 8-step methods, of order 8, and symmetric, i.e., the relations $\alpha_i = \alpha_{k-i}$ and $\beta_i = \beta_{k-i}$ are satisfied. Therefore, we present the coefficients only for $i \leq k/2$.

These methods give approximations y_n to the solution of the differential equation. If also derivative approximations are needed, we get them by finite differences, e.g., for the 8th order methods of Table 7.1 we use

$$\dot{y}_n = \frac{1}{840h}\Big(672\,(y_{n+1}-y_{n-1}) - 168\,(y_{n+2}-y_{n-2}) + 32\,(y_{n+3}-y_{n-3}) - 3\,(y_{n+4}-y_{n-4})\Big). \tag{7.2}$$

We apply this method to the Kepler problem (I.2.2), once with eccentricity $e = 0$ and once with $e = 0.2$, and initial values (I.2.11), such that the period of the exact solution is 2π. Starting approximations are computed accurately with a high order Runge–Kutta method. We apply Method SY8 with many different step sizes ranging from $2\pi/30$ to $2\pi/95$, and we plot in Fig. 7.1 the maximum error of the total energy as a function of $2\pi/h$ (where h denotes the step size). We see that in general the error decreases with the step size, but there is an extremely large error for $h \approx 2\pi/60$. For $e \neq 0$, further peaks can be observed at integral multiples of 5 and 6. It is our aim to understand this behaviour.

Instabilities. We put $z = q_1 + iq_2$, so that the Kepler problem becomes

$$\ddot{z} = \psi(|z|)z, \qquad \psi(r) = -r^{-3},$$

and we choose initial values such that $z(t) = e^{it}$ is a circular motion (eccentricity $e = 0$). The numerical solution of (1.8) is therefore defined by the relation

$$\sum_{j=0}^{k} \alpha_j z_{n+j} = h^2 \sum_{j=0}^{k} \beta_j \psi(|z_{n+j}|) z_{n+j}. \tag{7.3}$$

Approximating $\psi(|z_{n+j}|)$ with $\psi(1) = -\omega^2$, we get a linear recurrence relation with characteristic polynomial

$$S(\omega h, \zeta) = \rho(\zeta) + \omega^2 h^2 \sigma(\zeta).$$

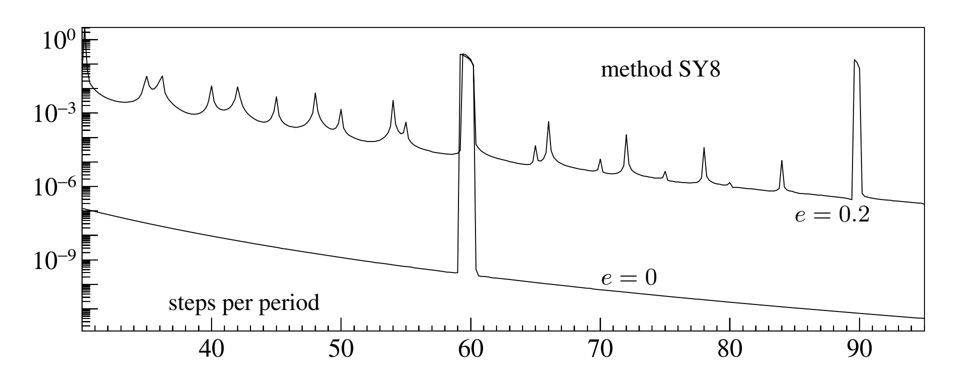


Fig. 7.1. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

The principal roots of $S(\omega h, \zeta) = 0$ satisfy $\zeta_1(\omega h) \approx e^{i\omega h}$ and $\zeta_2(\omega h) \approx e^{-i\omega h}$, and we have $|\zeta_j(\omega h)| = 1$ for all j and for sufficiently small h, because the method is symmetric (Exercise 2). As a consequence of $|\zeta_1(\omega h)| = 1$, the values $\widehat{z}_n := \zeta_1(\omega h)^n$ are not only a solution of the linear recurrence relation, but also of the nonlinear relation (7.3). Our aim is to study the stability of this numerical solution. We therefore consider a perturbed solution

$$z_n = \zeta_1(\omega h)^n \big(1 + u_n\big).$$

Using $|z_n| = 1 + \frac{1}{2}(u_n + \overline{u}_n) + \mathcal{O}(|u_n|^2)$ and neglecting the quadratic and higher order terms of $|u_n|$ in the relation (7.3), we get

$$\sum_{j=0}^{k} (\alpha_j + \omega^2 h^2 \beta_j) \zeta_1(\omega h)^j u_{n+j} = \frac{h^2}{2} \psi'(1) \sum_{j=0}^{k} \beta_j \zeta_1(\omega h)^j \big(u_{n+j} + \overline{u}_{n+j}\big).$$

Considering also the complex conjugate of this relation, and eliminating $\overline{u}_{n+j}$, we obtain a linear recurrence relation for u_n with characteristic polynomial

$$S(\omega h, \zeta_1(\omega h)\zeta) \cdot S(\omega h, \zeta_1(\omega h)^{-1}\zeta) + \mathcal{O}(h^2). \tag{7.4}$$

For small h, its zeros are close to $\zeta_1(\omega h)^{-1}\zeta_j$ and $\zeta_1(\omega h)\zeta_l$. If two of these zeros collapse, the $\mathcal{O}(h^2)$ terms in (7.4) can produce a root of modulus larger than one, so that instability occurs. This is the case, if two roots ζ_j, ζ_l of $\rho(\zeta) = 0$ satisfy $\zeta_j \zeta_l^{-1} \approx \zeta_1^2 \approx e^{2i\omega h}$, or

$$\theta_j - \theta_l = \frac{4\pi}{N}, \tag{7.5}$$

where $\zeta_j = e^{i\theta_j}$ and $h = 2\pi/N$.

For the Method SY8 of Table 7.1, the spurious zeros of $\rho(\zeta)$ have arguments $\pm 4\pi/5$, $\pm 2\pi/5$, and $\pm 2\pi/6$. With $\theta_j = 2\pi/5$ and $\theta_l = 2\pi/6$, the condition (7.5) gives $N = 60$ as a candidate for instability. This explains the experiment of Fig. 7.1 for $e = 0$. A study of the stability of orbits with eccentricity $e \neq 0$ (see Quinlan

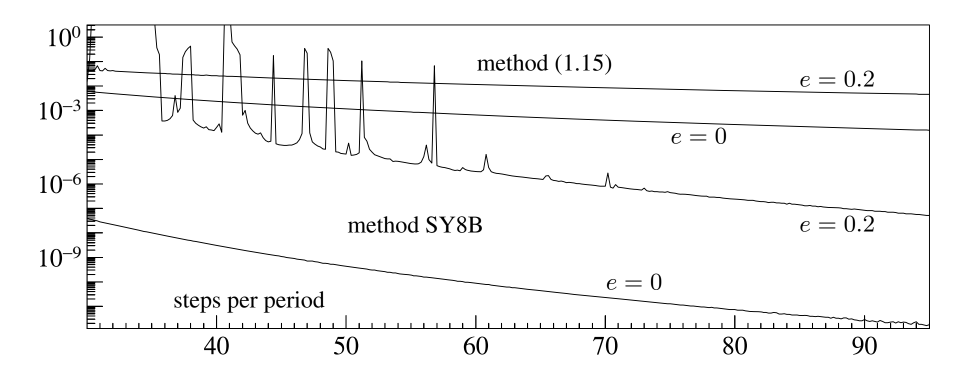


Fig. 7.2. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

1999) shows that instabilities can also occur when $4\pi/N$ is replaced with $2q\pi/N$ ($q = 2, 3, \ldots$) in the relation (7.5).

To avoid these instabilities as far as possible, Quinlan (1999) constructed symmetric multistep methods, where the spurious roots of $\rho(\zeta) = 0$ are well spread out on the unit circle and far from $\zeta = 1$. As a result he proposes Method SY8B of Table 7.1. The same experiment as above yields the results of Fig. 7.2. The ρ-polynomial of Method SY8B is

$$\rho(\zeta) = (\zeta - 1)^2\big(\zeta^6 + 2\zeta^5 + 3\zeta^4 + 3.5\zeta^3 + 3\zeta^2 + 2\zeta + 1\big),$$

and the θ_j of the spurious roots are $\pm 2\pi/2.278$, $\pm 2\pi/3.353$, and $\pm 2\pi/4.678$. The condition (7.5) is satisfied only for $N \leq 23.67$, which implies that no instability occurs for $e = 0$ in the region of the experiment of Fig. 7.2.

To illustrate the importance of high order methods, we included in Fig. 7.2 the results of the second order partitioned multistep method (1.15).

XV.7.2 Extension to Variable Step Sizes

Variable step size multistep methods for second order differential equations $\ddot{y} = f(y)$ are of the form

$$\sum_{j=0}^{k} \alpha_j(h_n, \ldots, h_{n+k-1})\, y_{n+j} \;=\; h_{n+k-1}^2 \sum_{j=0}^{k} \beta_j(h_n, \ldots, h_{n+k-1})\, f\big(y_{n+j}\big),$$

where the coefficients α_j and β_j are allowed to depend on the step sizes $h_n, \ldots$, h_{n+k-1}, more precisely, on the ratios $h_{n+1}/h_n, \ldots, h_{n+k-1}/h_{n+k-2}$. They yield approximations y_n to $y(t_n)$ on a variable grid given by $t_{n+1} = t_n + h_n$. Such a method is of *order* r (cf. formula (1.9)), if

$$\sum_{j=0}^{k} \alpha_j(h_n, \ldots, h_{n+k-1})\, y(t_{n+j}) \;=\; h_{n+k-1}^2 \sum_{j=0}^{k} \beta_j(h_n, \ldots, h_{n+k-1})\, \ddot{y}(t_{n+j}) \tag{7.6}$$

for all polynomials $y(t)$ of degree $\leq r+1$. It is *stable*, if the ρ-polynomial with coefficients $\alpha_j(h,\ldots,h)$ (constant step size) satisfies the stability condition of Sect. XV.1.2 (see Theorem III.5.7 of Hairer, Nørsett & Wanner (1993) and Cano & Durán (2003a)).

All methods of Sect. XV.7.1 can be extended to symmetric, variable step size integrators. This has been discovered by Cano & Durán (2003b). For clarity of notation we let $\widetilde\alpha_j, \widetilde\beta_j$ ($j=0,\ldots,k$) be the coefficients of such a fixed step size method. Cano & Durán propose putting

$$\beta_j(h_n,\ldots,h_{n+k-1}) = \frac{h_n}{h_{n+k-1}}\,\widetilde\beta_j, \tag{7.7}$$

and to determine $\alpha_j(h_n,\ldots,h_{n+k-1})$ such that symmetry and order $k-2$ (for arbitrary step sizes) are achieved. We also suppose (7.7), but we determine the coefficients $\alpha_j(h_n,\ldots,h_{n+k-1})$ such that (7.6) holds for all polynomials $y(t)$ of degree $\leq k$. This uniquely determines these coefficients whenever $h_n>0,\ldots$, $h_{n+k-1}>0$ (Vandermonde type system) and gives the following properties.

Lemma 7.1. *For even k, let $(\widetilde\alpha_j, \widetilde\beta_j)$ define a symmetric, stable k-step method (1.8) of order k, and consider the variable step size method given by (7.7) and $\alpha_j(h_n,\ldots,h_{n+k-1})$ such that (7.6) holds for all polynomials y satisfying $\deg y\leq k$. This method extends the fixed step size formula, i.e.,*

$$\alpha_j(h,\ldots,h) = \widetilde\alpha_j, \qquad \beta_j(h,\ldots,h) = \widetilde\beta_j, \tag{7.8}$$

it satisfies the symmetry relations

$$\begin{aligned} \alpha_j(h_n,\ldots,h_{n+k-1}) &= \alpha_{k-j}(h_{n+k-1},\ldots,h_n) \\ h_{n+k-1}^2\,\beta_j(h_n,\ldots,h_{n+k-1}) &= h_n^2\,\beta_{k-j}(h_{n+k-1},\ldots,h_n), \end{aligned} \tag{7.9}$$

and it is of order $k-1$ for arbitrary step sizes. Moreover, it behaves like a method of order k, if $h_{n+1}=h_n\big(1+\mathcal{O}(h_n)\big)$ uniformly in n.

Proof. The relation (7.8) for β_j follows at once from (7.7), and for α_j it is a consequence of the uniqueness of the solution of the linear system for the α_j.

The second condition of (7.9) follows directly from (7.7) and from the symmetry of the underlying fixed step size method ($\widetilde\beta_{k-j}=\widetilde\beta_j$ for all j). Inserting (7.7) into (7.6), replacing $y(t)$ with $y(t_{n+k}+t_n-t)$, and reversing the order of $h_n,\ldots,h_{n+k-1}$ yields

$$\sum_{j=0}^k \alpha_j(h_{n+k-1},\ldots,h_n)\,y(t_{n+k-j}) \;=\; h_n h_{n+k-1}\sum_{j=0}^k \widetilde\beta_j\,\ddot y(t_{n+k-j}).$$

Using $\widetilde\beta_{k-j}=\widetilde\beta_j$ this shows that $\alpha_{k-j}(h_{n+k-1},\ldots,h_n)$ satisfies exactly the same linear system as $\alpha_j(h_n,\ldots,h_{n+k-1})$, so that also the first relation of (7.9) is verified.

By definition, the variable step size method is at least of order $k-1$. Under the assumption $h_{n+1} = h_n\big(1+\mathcal{O}(h_n)\big)$ the defect in (7.6) is of the form

$$h_n^{k+1} D(h_n,\ldots,h_{n+k-1}) = h_n^{k+1} D(h_n,\ldots,h_n) + \mathcal{O}(h_n^{k+2})$$

for all sufficiently smooth $y(t)$. Since the constant step size method is of order k, the expression $D(h_n,\ldots,h_n)$ is of size $\mathcal{O}(h_n)$, so that we observe convergence of order k. □

The symmetry relation (7.9) has the following interpretation: if the approximations $y_n,\ldots,y_{n+k-1}$ used with step sizes $h_n,\ldots,h_{n+k-1}$ yield y_{n+k}, then the values $y_{n+k},\ldots,y_{n+1}$ applied with $h_{n+k-1},\ldots,h_n$ yield y_n as a result (since the coefficients α_j and β_j only depend on step size ratios and the multistep formula only on h_{n+k-1}^2, the same result is obtained with $-h_{n+k-1},\ldots,-h_n$). This is the analogue of the definition of symmetry for one-step methods.

For obtaining a good long-time behaviour, the step sizes also have to be chosen in a symmetric and reversible way (see Sect. VIII.3). One possibility is to take step sizes

$$h_{n+k-1} = \frac{\varepsilon}{2}\Big(\sigma(y_{n+k-1}) + \sigma(y_{n+k})\Big), \tag{7.10}$$

where $\varepsilon > 0$, and $\sigma(y)$ is a given positive monitor function. This condition is an implicit equation for h_{n+k-1}, because y_{n+k} depends on h_{n+k-1}. It has to be solved iteratively. Notice, however, that for an explicit multistep formula no further force evaluations are necessary during this iteration. Such a choice of the step size guarantees that whenever h_{n+k-1} is chosen when stepping from $y_n,\ldots,y_{n+k-1}$ with $h_n,\ldots,h_{n+k-2}$ to y_{n+k}, the step size h_n is chosen when stepping backwards from $y_{n+k},\ldots,y_{n+1}$ with $h_{n+k-1},\ldots,h_{n+1}$ to y_n.

Implementation. For given initial values $y_0, \dot y_0$, the starting approximations y_1, $\ldots, y_{k-1}$ should be computed accurately (for example, by a high-order Runge–Kutta method) with step sizes satisfying (7.10). The solution of the scalar nonlinear equation (7.10) has to be done carefully in order to reduce the overhead of the method. In our code we use $h_{n+k-1} := h_{n+k-2}^2/h_{n+k-3}$ as predictor, and we apply modified Newton iterations with the derivative approximated by finite differences.

The coefficients $\alpha_j(h_n,\ldots,h_{n+k-1})$ have to be computed anew in every iteration. We use the basis

$$p_i(t) = \prod_{j=0}^{i-1}(t-t_{n+j}), \qquad i=0,\ldots,k$$

for the polynomials of degree $\le k$ in (7.6). This leads to a linear triangular system for $\alpha_0,\ldots,\alpha_k$. As noticed by Cano & Durán (2003b), the coefficients $p_i(t_j)$ and $\ddot p_i(t_j)$ can be obtained efficiently from the recurrence relations

$$\begin{aligned}
p_0(t) &= 1, &\quad p_{i+1}(t) &= (t-t_i)p_i(t)\\
\dot p_0(t) &= 0, &\quad \dot p_{i+1}(t) &= (t-t_i)\dot p_i(t) + p_i(t)\\
\ddot p_0(t) &= 0, &\quad \ddot p_{i+1}(t) &= (t-t_i)\ddot p_i(t) + 2\dot p_i(t).
\end{aligned}$$

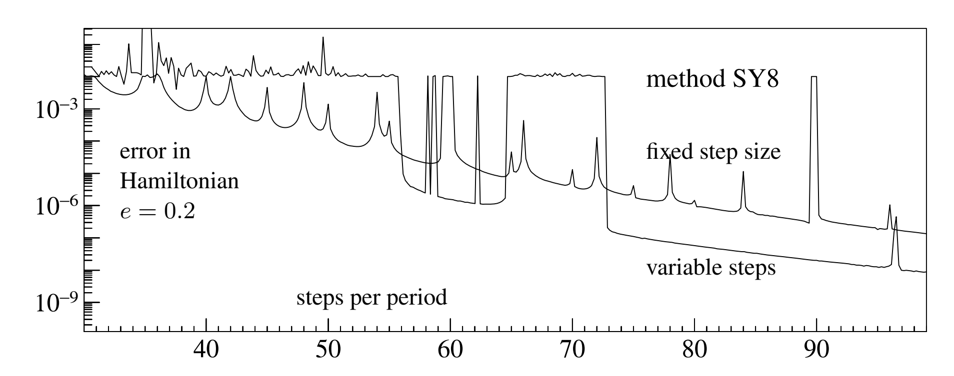


Fig. 7.3. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

During the iterations for the solution of the nonlinear equation (7.10) only the values of $p_i(t_{n+k})$ have to be updated.

Numerical Experiment. We repeat the experiment of Fig. 7.1 with the method SY8, but this time in the variable step size version and with $\sigma(y) = \|y\|^2$ as step size monitor. We have computed 2500 periods of the Kepler problem with eccentricity $e = 0.2$, and we have plotted in Fig. 7.3 the maximal error in the Hamiltonian as a function of the number of steps per period (for a comparison we have also included the result of the fixed step size implementation). Similar to (7.2) we use approximations $\dot{y}_n$ that are the derivative of the interpolation polynomial passing through $y_n, y_{n\pm1}, y_{n\pm2}, \ldots$ such that the correct order is obtained. The computation is stopped when the error exceeds 10^{-2}.

As expected, the error is smaller for the variable step size version, and it is seen that the peaks due to numerical resonances are now much less although they are not completely removed. For large step sizes, the performance deteriorates, but this is not a serious problem, because these methods are recommended only for high accuracy computations.

It should be remarked that the overhead, due to the computation of the coefficients α_j and the solution of the nonlinear equation (7.10), is rather high. Therefore, the use of variable step sizes is recommended only when force evaluations $f(y)$ are expensive or when constant step sizes are not appropriate. Cano & Durán (2003b) report an excellent performance of symmetric, variable step size multistep methods for computations of the outer solar system.

> Despite the resonances and instabilities, then, symmetric methods can still be a better choice than Störmer methods for long integrations of planetary orbits provided that the user is aware of the dangers.
> (G.D. Quinlan 1999)

XV.8 Multi-Value or General Linear Methods

General linear methods is a class of integration methods that covers Runge–Kutta as well as multistep methods. It is therefore of interest to study which of the results on the long-time behaviour can be extended.

So-called multi-value or general linear methods are defined by $Y_{n+1} = G_h(Y_n)$, where

$$\begin{aligned} Y_{n+1} &= DY_n + hBf(U_{n+1}) \\ U_{n+1} &= CY_n + hAf(U_{n+1}) \end{aligned} \tag{8.1}$$

with $f(U_{n+1}) = \big(f(u^1_{n+1}),\dots,f(u^s_{n+1})\big)^T$ for $U_{n+1} = \big(u^1_{n+1},\dots,u^s_{n+1}\big)^T$, and $Y_n = (y^1_n,\dots,y^k_n)$. We use a sloppy notation in the sense that the matrices $D, B, \dots$ should be replaced with $D\otimes I, B\otimes I, \dots$. For a computation, a starting procedure S_h and a finishing procedure F_h, which extracts the numerical approximation y_n from Y_n, have to be added (see Fig. 8.1). We assume the existence of a vector e such that with $\mathbb{1} = (1,\dots,1)^T$

$$De = e, \qquad Ce = \mathbb{1} \tag{8.2}$$

holds (preconsistency conditions). The vector Y_n is then an approximation to $ey(t_n)$ (more precisely to $e\otimes y(t_n)$).

For Runge–Kutta methods, $D = (1)$ is the one-dimensional identity, $B = (b_1,\dots,b_s)$, $C = \mathbb{1}$, and A is the usual Runge-Kutta matrix. For multistep methods, we have $Y_n = (y_{n+k-1},\dots,y_n)^T$, and D is the $k\times k$ matrix with characteristic polynomial $\rho(\zeta)$ as in (2.1). For a detailed treatment of general linear methods we refer the reader to Chap. 4 of the monograph of Butcher (1987), and to Chap. III.8 of Hairer, Nørsett & Wanner (1993).

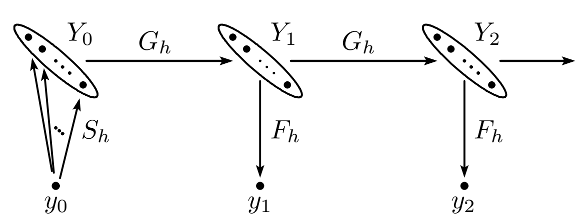
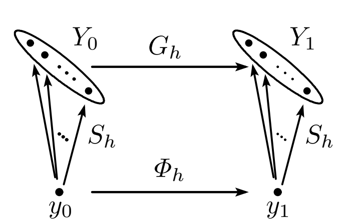


Fig. 8.1. Illustration of a multi-value method $Y_{n+1} = G_h(Y_n)$ with starting procedure S_h and finishing procedure F_h

XV.8.1 Underlying One-Step Method and Backward Error Analysis

In analogy to multistep methods, a method (8.1) is called strictly stable, if all eigenvalues of D are inside the unit circle with the exception of the single eigenvalue $\zeta = 1$. An extension of Kirchgraber's result (Theorem 2.1) to strictly stable general linear methods is given by Stoffer (1993).

Theorem 8.1. *Consider a strictly stable general linear method $Y_{n+1} = G_h(Y_n)$, and a finishing procedure $y_n = F_h(Y_n) = d^T Y_n + \mathcal{O}(h)$. Assume that (8.2) and $d^T e = 1$ hold.*



(i) Then there exist a unique one-step method $\Phi_h(y)$ and a unique starting procedure $S_h(y)$ such that $G_h \circ S_h = S_h \circ \Phi_h$ and $F_h \circ S_h = Id$ hold.

(ii) The manifold $\mathcal{M}_h = \{S_h(y)\,;\, y \in \mathbb{R}^d\}$ is invariant under G_h, and it is exponentially attractive.

Proof. Since the method is strictly stable, there exists a matrix T such that

$$T^{-1}DT = \begin{pmatrix} 1 & 0 \\ 0 & D_0 \end{pmatrix} \qquad \text{with} \quad \|D_0\| < 1,$$

and $Te_1 = e$ (where $e_1 = (1, 0, \ldots, 0)^T$). The proof closely follows that of Theorem 2.1. With the transformation $(\xi_n, \eta_n)^T = Z_n = T^{-1}Y_n$, the general linear method (8.1) becomes

$$\begin{pmatrix} \xi_{n+1} \\ \eta_{n+1} \end{pmatrix} = \begin{pmatrix} \xi_n \\ D_0 \eta_n \end{pmatrix} + hT^{-1}Bf\big(U_{n+1}\big). \tag{8.3}$$

with $U_{n+1} = CTZ_n + hAf(U_{n+1})$. As before, Theorem XII.3.1 can be applied and yields the existence of an attractive manifold $\mathcal{N}_h = \{(\xi, s(\xi))\,;\, \xi \in \mathbb{R}^d\}$, which is invariant under the mapping (8.3). We now invert the restriction of F_h onto the manifold $\mathcal{N}_h$. Due to $d^T e = 1$ and $Te_1 = e$, we have for $Z = Z(\xi) = \big(\xi, s(\xi)\big)^T$ that

$$y = F_h\big(TZ(\xi)\big) = d^T TZ(\xi) + \ldots = \xi + g(\xi), \tag{8.4}$$

where $g(\xi)$ is Lipschitz continuous with constant $\mathcal{O}(h)$. By the Banach fixed-point theorem the equation (8.4) has a unique solution $\xi = r(y)$. Putting

$$S_h(y) = TZ\big(r(y)\big) = T \begin{pmatrix} r(y) \\ s\big(r(y)\big) \end{pmatrix},$$

we have found the unique starting procedure satisfying $F_h \circ S_h = Id$ and $T^{-1}S_h(y) \in \mathcal{N}_h$. We finally define $\Phi_h = F_h \circ G_h \circ S_h$ and $\mathcal{M}_h = \{TZ\,;\, Z \in \mathcal{N}_h\}$, so that all statements of the theorem are verified. □

It is our aim to extend the concept of an underlying one-step method to nearly all (including weakly stable) general linear methods.

Theorem 8.2. *Consider a general linear method (8.1), and assume that $\zeta = 1$ is a single eigenvalue of the propagation matrix D. Furthermore, let $G_h(Y)$ and $F_h(Y) = d^T Y + \ldots$ have expansions in powers of h, and assume that (8.2) and $d^T e = 1$ hold. Then there exist a unique formal one-step method*

$$\Phi_h(y) = y + hd_1(y) + h^2 d_2(y) + \ldots$$

and a unique formal starting procedure

$$S_h(y) = ey + hS_1(y) + h^2 S_2(y) + \ldots,$$

such that formally $G_h \circ S_h = S_h \circ \Phi_h$ *and* $F_h \circ S_h = Id$ *hold.*

Proof. Expanding $S_h\big(\Phi_h(y)\big)$ and $G_h\big(S_h(y)\big)$ into powers of h, a comparison of the coefficients yields

$$ed_j(y) + (I - D)S_j(y) = \ldots \ , \tag{8.5}$$

where a right-hand side depends on known functions and on $d_i(y), S_i(y)$ with $i < j$. Similarly, the condition $F_h\big(S_y(y)\big) = y$ leads to

$$d^T S_j(y) = \ldots \ . \tag{8.6}$$

Due to the fact that $\zeta = 1$ is a single eigenvalue of D, and that $d^T e \neq 0$, the system (8.5)-(8.6) uniquely determines $d_j(y)$ and $S_j(y)$. □

Backward Error Analysis for Smooth Numerical Solutions. The formal analysis of Chap. IX can be directly applied to the underlying one-step method of Theorem 8.2. This yields a modified differential equation, but only for the smooth numerical solution (cf. Sect. XV.3.1). Notice that this modified equation depends on the choice of the finishing procedure F_h.

XV.8.2 Symplecticity and Symmetry

Before giving a precise meaning to the symplecticity and symmetry of general linear methods, we establish the following lemma.

Lemma 8.3. *For a general linear method* $Y_{n+1} = G_h(Y_n)$ *we consider two different finishing procedures* $y_n = F_h(Y_n)$ *and* $\widehat{y}_n = \widehat{F}_h(Y_n)$ *:*

$$\begin{array}{ccccccc}
\widehat{y}_0 & \xrightarrow{\widehat{\Phi}_h} & \widehat{y}_1 & \xrightarrow{\widehat{\Phi}_h} & \widehat{y}_2 & \xrightarrow{\widehat{\Phi}_h} & \ldots \\
{\scriptstyle \widehat{S}_h}\big\downarrow\big\uparrow{\scriptstyle \widehat{F}_h} & & \big\uparrow{\scriptstyle \widehat{F}_h} & & \big\uparrow{\scriptstyle \widehat{F}_h} & & \\
Y_0 & \xrightarrow{G_h} & Y_1 & \xrightarrow{G_h} & Y_2 & \xrightarrow{G_h} & \ldots \\
{\scriptstyle S_h}\big\uparrow\big\downarrow{\scriptstyle F_h} & & \big\downarrow{\scriptstyle F_h} & & \big\downarrow{\scriptstyle F_h} & & \\
y_0 & \xrightarrow{\Phi_h} & y_1 & \xrightarrow{\Phi_h} & y_2 & \xrightarrow{\Phi_h} & \ldots
\end{array}$$

The two corresponding one-step methods $\Phi_h(y)$ *and* $\widehat{\Phi}_h(y)$ *(given by Theorem 8.2) are then conjugate to each other, i.e.,*

$$\alpha_h^{-1} \circ \Phi_h \circ \alpha_h = \widehat{\Phi}_h \qquad \textit{with} \qquad \alpha_h = F_h \circ \widehat{S}_h. \tag{8.7}$$

Proof. The equations involving the underlying one-step methods or the starting procedures have to be understood in the sense of formal series. By Theorem 8.2 we have $S_h(y) = ey + \mathcal{O}(h)$ and also $\widehat{S}_h(y) = ey + \mathcal{O}(h)$. It thus follows from $F_h \circ S_h = Id$ that $\alpha_h(y)$ is $\mathcal{O}(h)$-close to the identity and therefore invertible. □

The transformation α_h in the phase space is $\mathcal{O}(h)$-close to the identity. The relation $\alpha_h^{-1} \circ \Phi_h^n \circ \alpha_h = \widehat{\Phi}_h^n$, which is a consequence of (8.7), therefore implies that the numerical solutions of Φ_h and $\widehat{\Phi}_h$ remain $\mathcal{O}(h)$-close for all times. This means that the long-time behaviour of both methods is exactly the same.

Consequently, for a given general linear method G_h, it is sufficient to require symplecticity or symmetry for *one* finishing procedure only.

Definition 8.4 (Symplecticity). A general linear method G_h is called *symplectic* if there exists a finishing procedure F_h such that the underlying one-step method Φ_h of Theorem 8.2 is symplectic, i.e., $\Phi_h'(y)^T J \,\Phi_h'(y) = J$ in the sense of formal series.

The study of symplecticity of linear multistep methods (Sect. XV.4.1) was rather disappointing. We could not find one linear multistep method whose underlying one-step method is symplectic. For general linear methods, some necessary conditions for the symplecticity of the underlying one-step method are known which are hard to satisfy (Hairer & Leone 1998). For the moment, no symplectic general linear method (not equivalent to a one-step method) is known, and we conjecture that such a method does not exist, even in the class of partitioned general linear methods (treating the p and q variables by different methods).

After the disappointing non-existence conjecture of symplectic multi-value methods, we turn our attention to symmetric methods. We know from the previous chapters that for reversible Hamiltonian systems, the long time behaviour of symmetric one-step methods can be as good as that for symplectic methods. There are several definitions of symmetric general linear methods in the literature. However, they are either tailored to very special situations (e.g., Hairer, Nørsett & Wanner 1993), or they do not allow the proof of results that are expected to hold for symmetric methods.

Definition 8.5 (Symmetry). A general linear method G_h is called *symmetric* if there exists a finishing procedure F_h such that the underlying one-step method Φ_h of Theorem 8.2 is symmetric, i.e., $\Phi_{-h}(y) = \Phi_h^{-1}(y)$ in the sense of formal series.

Example 8.6. Consider the trapezoidal method in the role of G_h and the explicit Euler method with step size $-\gamma h$ as finishing procedure:

$$\begin{aligned} G_h: \qquad & Y_{n+1} = Y_n + \frac{h}{2}\Big(f(Y_n) + f(Y_{n+1})\Big) \\ F_h: \qquad & y_{n+1} = Y_{n+1} - \gamma h f(Y_{n+1}) \end{aligned}$$

The corresponding starting procedure and underlying one-step methods are then the implicit Euler method and the following 2-stage Runge–Kutta method:

$$S_h: \qquad Y_n = y_n + \gamma h f(Y_n)$$

$$\Phi_h: \qquad \text{Runge–Kutta method} \qquad \begin{array}{c|cc} \gamma & \gamma & \\ 1+\gamma & 1/2+\gamma & 1/2 \\ \hline & 1/2+\gamma & 1/2-\gamma \end{array}$$

The method Φ_h is symmetric only for $\gamma = 0$, for $\gamma = 1/2$, and for $\gamma = -1/2$. This example demonstrates that the symmetry of the underlying one-step method strongly depends on the finishing procedure.

On the other hand, this example shows that the 2-stage Runge–Kutta method is symmetric in the sense of Definition 8.5 for all γ (because it is conjugate to the trapezoidal rule). It is not symmetric according to the definition of Chap. V.

A Useful Criterion for Symmetry. Definition 8.5 is rather impractical for verifying the symmetry of a given general linear method. We give here algebraic conditions for the coefficients A, B, C, D of a general linear method (8.1), which are sufficient for the method to be symmetric. We assume that the finishing procedure $y_{n+1} = F_h(Y_{n+1})$ is given by

$$y_{n+1} = \widetilde{D} Y_{n+1} + h\widetilde{B} f(V_{n+1}), \qquad V_{n+1} = \widetilde{C} Y_{n+1} + h\widetilde{A} f(V_{n+1}), \tag{8.8}$$

in complete analogy to method (8.1).

Lemma 8.7 (Adjoint Method). *Let $Y_{n+1} = G_h(Y_n)$ be the general linear method given by A, B, C, D (with invertible D), $y_{n+1} = F_h(Y_{n+1})$ the finishing procedure given by $\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}$, and denote by Φ_h its underlying one-step method. Then, the underlying one-step method of*

$$\begin{array}{lllll} G_h^*: & A^* = CD^{-1}B - A, & B^* = D^{-1}B, & C^* = CD^{-1}, & D^* = D^{-1} \\ F_h^*: & \widetilde{A}^* = -\widetilde{A}, & \widetilde{B}^* = -\widetilde{B}, & \widetilde{C}^* = \widetilde{C}, & \widetilde{D}^* = \widetilde{D} \end{array}$$

is the adjoint method $\Phi_h^* = \Phi_{-h}^{-1}$ of Φ_h.

Proof. Substituting $h \leftrightarrow -h$ and $Y_{n+1} \leftrightarrow Y_n$ in (8.1) yields

$$U_{n+1} = CY_{n+1} - hAf(U_{n+1}), \qquad Y_n = DY_{n+1} - hBf(U_{n+1}).$$

Extracting Y_{n+1} from the second relation and inserting it into the first gives

$$\begin{aligned} U_{n+1} &= CD^{-1}Y_n + h(CD^{-1}B - A)f(U_{n+1}) \\ Y_{n+1} &= D^{-1}Y_n + hD^{-1}Bf(U_{n+1}), \end{aligned}$$

which is exactly method G_h^*. The same replacements in the finishing procedure

$$V_{n+1} = \widetilde{C} Y_n - h\widetilde{A} f(V_{n+1}), \qquad y_n = \widetilde{D} Y_n - h\widetilde{B} f(V_{n+1})$$

and in the diagram of Theorem 8.2 prove the statement. □

Theorem 8.8. *If there exist an invertible matrix Q (satisfying $Qe = e$ with e given by (8.2)) and a permutation matrix P such that*

$$\begin{aligned} P^{-1}AP &= CD^{-1}B - A, & Q^{-1}BP &= D^{-1}B, \\ P^{-1}CQ &= CD^{-1}, & Q^{-1}DQ &= D^{-1}, \end{aligned} \tag{8.9}$$

then the general linear method (8.1) is symmetric.

Proof. We consider the change of variables $Y_n = Q\,\widehat{Y}_n$, $U_n = P\,\widehat{U}_n$ in the method (8.1). Since P is a permutation matrix, we have $f(PU) = Pf(U)$, so that the method becomes

$$P\widehat{U}_{n+1} = CQ\widehat{Y}_n + hAPf(\widehat{U}_{n+1}), \qquad Q\widehat{Y}_{n+1} = DQ\widehat{Y}_n + hBPf(\widehat{U}_{n+1}).$$

The assumption (8.9) implies that this method is the same as the adjoint method of Lemma 8.7. Taking a finishing procedure F_h in such a way that $y_{n+1} = F_h(Q\,\widehat{Y}_{n+1})$ is identical to the finishing procedure $y_{n+1} = F_h^*(\widehat{Y}_{n+1})$ of the adjoint method (i.e., $\widetilde{B} = 0$ and $\widetilde{D}$ such that $\widetilde{D}Q = \widetilde{D}$), we obtain $\Phi_h^* = \Phi_h$. This proves the statement. □

The sufficient condition of Theorem 8.8 reduces to the known criteria for classical methods. Let us give some examples:

- For Runge–Kutta methods we have $D = (1)$, $B = b^T$ a row vector, and $C = 1\!\!1$. With $Q = (1)$ and P the permutation matrix that inverts the elements of a vector, we get
$$b^T P = b^T, \qquad PAP = 1\!\!1 b^T - A,$$
which is the same (V.2.4).
- Multistep methods in their form as general linear methods (Sect. XV.8) satisfy the condition of Theorem 8.8 if
$$\alpha_i = -\alpha_{k-i}, \qquad \beta_i = \beta_{k-i}. \tag{8.10}$$
One can take for P and Q the permutation matrices (inverting the elements of a vector) of dimension $k+1$ and k, respectively.

XV.8.3 Growth Parameters

For a rigorous study of the long-time behaviour of general linear methods it is not sufficient to investigate smooth numerical solutions. One has to get bounds on the parasitic solution components, which are present when one considers the general linear method without any starting and finishing procedure. This is certainly difficult, as it is for multistep methods (1.1). We restrict here our analysis to the linearized parasitic modified equation.

The eigenvalues of the matrix D in (8.1) will play the role of the zeros of $\rho(\zeta)$ in (1.1). We denote them by $\zeta_1 = 1$ and $\zeta_2, \ldots, \zeta_k$, and we assume that they are simple

and of modulus one. Motivated by the analysis for multistep methods we write the approximations Y_n as

$$Y_n = Y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n Z_\ell(nh) \tag{8.11}$$

with smooth functions $Y(t)$ and $Z_\ell(t)$. The index set $\mathcal{I}^*$ has the same meaning as in Sect. XV.3.2. We insert (8.11) into (8.1) and compare coefficients of ζ_ℓ^n. This gives with $t = nh$

$$\begin{aligned} Y(t+h) &= DY(t) + hBf\big(CY(t)\big) + \mathcal{O}(h^2) \\ \zeta_\ell Z_\ell(t+h) &= DZ_\ell(t) + hBf'\big(CY(t)\big)CZ_\ell(t) + \mathcal{O}(h^2). \end{aligned} \tag{8.12}$$

To get an amenable form of the modified equations we write the vectors $Y(t), Z_\ell(t)$ in the basis of eigenvectors of D, which we denote by $w_1 = e$ and $w_2, \ldots, w_k$:

$$Y(t) = \sum_{j=1}^{k} y_j(t)\, w_j, \qquad Z_\ell(t) = \sum_{j=1}^{k} z_{\ell,j}(t)\, w_j.$$

Inserted into (8.12) and expanded into a series of h yields

$$\dot y_1 = f(y_1) + \mathcal{O}(h),$$

and algebraic relations of the form $y_j(t) = \mathcal{O}(h)$ for $j \geq 2$. Similarly, we get algebraic relations for $z_{\ell,j}(t) = \mathcal{O}(h)$ if $j \neq \ell$, and the function $z_\ell(t) := z_{\ell,\ell}(t)$ satisfies

$$\dot z_\ell = \mu_\ell f'(y_1) z_\ell + \mathcal{O}(h) \qquad \text{with} \qquad \mu_\ell = \zeta_\ell^{-1} w_j^* BCw_j, \tag{8.13}$$

where w_j^* is the left eigenvector of D corresponding to the eigenvalue ζ_ℓ. This is in perfect analogy to the computations of Sect. XV.5.1.

This analysis can be extended straightforwardly to partitioned general linear methods, where different methods are applied to the components y and v of a partitioned differential equation. Unfortunately, we do not know of any results that would extend those of Sect. XV.6 to general linear methods.

XV.9 Exercises

1. Let $\zeta_1(z)$ be the principal root of the characteristic equation $\rho(\zeta) - z\sigma(\zeta) = 0$. Prove that for irreducible multistep methods the condition $\zeta_1(-z)\zeta_1(z) \equiv 1$ (in a neighbourhood of $z = 0$) is equivalent to the symmetry of the method.
2. (Lambert & Watson 1976). Prove that stable, symmetric linear multistep methods (1.8) for second order differential equations, for which the polynomial $\rho(\zeta)$ has only simple zeros (with the exception of $\zeta = 1$), has a non-vanishing interval of periodicity, i.e., the roots $\zeta_i(z)$ of $\rho(\zeta) - z^2\sigma(\zeta) = 0$ satisfy $|\zeta_i(iy)| = 1$ for sufficiently small real y.
   Hint. Simple roots cannot leave the unit circle under small perturbations of y.

3. Consider a symmetric, s-stable multistep method (1.8). If it is irreducible (no common factors of $\rho(\zeta)$ and $\sigma(\zeta)$), then k is even. Hence $\rho(-1) \neq 0$.
4. Using Theorem XII.3.2, prove that the underlying one-step method of a strictly stable rth order linear multistep method has order r.
5. (Dahlquist 1959). Consider the linear problem $\dot{y} = \lambda y$ and apply a symmetric linear multistep method (1.1) as in Example 2.2. Prove that for $t = nh$ and $h \to 0$,
$$\zeta_j^n(\lambda h) \approx \zeta_j^n e^{\mu_j \lambda t},$$
where μ_j is the growth parameter.
6. Consider a general linear method (8.1). If there exist an invertible symmetric matrix G and a diagonal matrix Λ such that
$$M = \begin{pmatrix} D^T G D - G & D^T G B - C^T \Lambda \\ B^T G D - \Lambda C & B^T G B - A^T \Lambda - \Lambda A \end{pmatrix} = 0, \tag{9.1}$$
then the method is G-symplectic.
Hint. Adapt the proof of Burrage & Butcher for B-stability (see Hairer & Wanner (1996), page 358).
7. A Runge–Kutta method can be considered as a general linear method with $D = (1)$, $C = \mathbb{1}$. Prove that the condition (9.1) is equivalent to the symplecticity condition of Chap. VI.
8. Extend the definition of G-symplecticity to partitioned general linear methods, and prove that the condition
$$M = \begin{pmatrix} D^T G \widehat{D} - G & D^T G \widehat{B} - C^T \Lambda \\ B^T G \widehat{D} - \Lambda \widehat{C} & B^T G \widehat{B} - A^T \Lambda - \Lambda \widehat{A} \end{pmatrix} = 0 \tag{9.2}$$
implies that the method is G-symplectic.
9. Construct general linear methods of order $r > 2$, for which all growth parameters are positive. Find such methods, which have a smaller degree of implicitness than symmetric one-step methods of the same order.
10. Write a Maple program that checks the coefficients of Table 7.1. After defining `rho:=`$\rho(z)$, use the instructions
```
> sigma := taylor(rho/(log(z)*log(z)),z=1,8);
> factor(expand(convert(sigma,polynom)));
```
11. Construct partitioned general linear methods which are symmetric, explicit, of high order, and for which the matrices D and $\widehat{D}$ have distinct eigenvalues (with the exception of 1). Compared to multistep methods, smaller dimensions of the matrices D and $\widehat{D}$ are possible.

Bibliography

R. Abraham & J.E. Marsden, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings Publishing Company, Reading, Massachusetts, 1978. *[XIV.3]*

L. Abia & J.M. Sanz-Serna, *Partitioned Runge–Kutta methods for separable Hamiltonian problems*, Math. Comput. 60 (1993) 617–634. *[VI.7], [IX.10]*

M.J. Ablowitz & J.F. Ladik, *A nonlinear difference scheme and inverse scattering*, Studies in Appl. Math. 55 (1976) 213–229. *[VII.4]*

M.P. Allen & D.J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987. *[I.4]*

H.C. Andersen, *Rattle: a "velocity" version of the Shake algorithm for molecular dynamics calculations*, J. Comput. Phys. 52 (1983) 24–34. *[VII.1]*

V.I. Arnold, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Russian Math. Surveys 18 (1963) 85–191. *[I.1]*

V.I. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaites*, Ann. Inst. Fourier 16 (1966) 319–361. *[VI.9]*

V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978, second edition 1989. *[VI.1], [VII.2], [VII.5], [X.1], [X.7]*

V.I. Arnold, V.V. Kozlov & A.I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer, Berlin, 1997. *[X.1]*

U. Ascher & S. Reich, *On some difficulties in integrating highly oscillatory Hamiltonian systems*, in Computational Molecular Dynamics, Lect. Notes Comput. Sci. Eng. 4, Springer, Berlin, 1999, 281–296. *[V.4]*

A. Aubry & P. Chartier, *Pseudo-symplectic Runge–Kutta methods*, BIT 38 (1998) 439–461. *[X.7]*

H.F. Baker, *Alternants and continuous groups*, Proc. of London Math. Soc. 3 (1905) 24–47. *[III.4]*

M.H. Beck, A. Jäckle, G.A. Worth & H.-D. Meyer, *The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propagating wavepackets*, Phys. Reports 324 (2000) 1–105. *[IV.9], [VII.6]*

G. Benettin, A.M. Cherubini & F. Fassò, *A changing-chart symplectic algorithm for rigid bodies and other Hamiltonian systems on manifolds*, SIAM J. Sci. Comput. 23 (2001) 1189–1203. *[VII.4]*

G. Benettin, L. Galgani & A. Giorgilli, *Poincaré's non-existence theorem and classical perturbation theory for nearly integrable Hamiltonian systems*, Advances in nonlinear dynamics and stochastic processes (Florence, 1985) World Sci. Publishing, Singapore, 1985, 1–22. *[X.2]*

G. Benettin, L. Galgani & A. Giorgilli, *Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. Part I*, Comm. Math. Phys. 113 (1987) 87–103. *[XIII.6]*

G. Benettin, L. Galgani & A. Giorgilli, *Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. II*, Commun. Math. Phys. 121 (1989) 557–601. *[XIII.9]*

G. Benettin, L. Galgani, A. Giorgilli & J.-M. Strelcyn, *A proof of Kolmogorov's theorem on invariant tori using canonical transformations defined by the Lie method*, Il Nuovo Cimento 79B (1984) 201–223. *[X.5]*

G. Benettin & A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys. 74 (1994) 1117–1143. *[IX.3], [IX.7], [IX.8]*

B.J. Berne, *Molecular dynamics in systems with multiple time scales: reference system propagator algorithms*, in Computational Molecular Dynamics: Challenges, Methods, Ideas (P. Deuflhard et al., eds.), Springer, Berlin 1999, 297–318. *[XIII.1]*

Joh. Bernoulli, *Problème inverse des forces centrales, extrait de la réponse de Monsieur Bernoulli à Monsieur Herman*, Mém. de l'Acad. R. des Sciences de Paris (1710) p. 521, *Opera Omnia* I, p. 470-480. *[I.2]*

M. Berry, *Histories of adiabatic quantum transitions*, Proc. Royal Soc. London A 429 (1990) 61–72. *[XIV.1]*

V. Betz & S. Teufel, *Precise coupling terms in adiabatic quantum evolution*, Ann. Henri Poincaré 6 (2005) 217–246. *[XIV.1]*

V. Betz & S. Teufel, *Precise coupling terms in adiabatic quantum evolution: the generic case*, Comm. Math. Phys., to appear (2005). *[XIV.1]*

J.J. Biesiadecki & R.D. Skeel, *Dangers of multiple time step methods*, J. Comput. Phys. 109 (1993) 318–328. *[I.4], [VIII.4], [XIII.1]*

G.D. Birkhoff, *Relativity and Modern Physics*, Harvard Univ. Press, Cambridge, Mass., 1923. *[I.6]*

G.D. Birkhoff, *Dynamical Systems*, AMS, Providence, R.I., 1927. *[X.2]*

S. Blanes, *High order numerical integrators for differential equations using composition and processing of low order methods*, Appl. Num. Math. (2001) 289–306. *[V.3]*

S. Blanes & F. Casas, *On the necessity of negative coefficients for operator splitting schemes of order higher than two*, Appl. Num. Math. 54 (2005) 23–37. *[III.3]*

S. Blanes, F. Casas & J. Ros, *Symplectic integrators with processing: a general study*, SIAM J. Sci. Comput. 21 (1999) 149–161. *[V.3]*

S. Blanes, F. Casas & J. Ros, *Improved high order integrators based on the Magnus expansion*, BIT 40 (2000a) 434–450. *[IV.7]*

S. Blanes, F. Casas & J. Ros, *Processing symplectic methods for near-integrable Hamiltonian systems*, Celestial Mech. Dynam. Astronom. 77 (2000b) 17–35. *[V.3]*

S. Blanes & P.C. Moan, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math. 142 (2002) 313–330. *[V.3]*

P.B. Bochev & C. Scovel, *On quadratic invariants and symplectic structure*, BIT 34 (1994) 337–345. *[VI.4], [XV.4]*

N. Bogolioubov & I. Mitropolski, *Les Méthodes Asymptotiques en Théorie des Oscillations Non Linéaires*, Gauthier-Villars, Paris, 1962. *[XII.2]*

N.N. Bogoliubov & Y.A. Mitropolsky, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Hindustan Publishing Corp., Delhi, 1961. *[XII.1]*

J.F. Bonnans & J. Laurent-Varin, *Computation of order conditions for symplectic partitioned Runge–Kutta schemes with application to optimal control*, Numer. Math., to appear (2006). *[VI.10]*

M. Born & V. Fock, *Beweis des Adiabatensatzes*, Zeitschr. f. Physik 51 (1928) 165–180. *[XIV.1], [XIV.4]*

F. Bornemann, *Homogenization in Time of Singularly Perturbed Mechanical Systems*, Springer LNM 1687 (1998). *[XIV.3]*

E. Bour, *L'intégration des équations différentielles de la mécanique analytique*, J. Math. Pures et Appliquées 20 (1855) 185–200. *[X.1]*

K.E. Brenan, S.L. Campbell & L.R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Appl. Math., SIAM, Philadelphia, 1996. *[IV.10]*

T.J. Bridges & S. Reich, *Computing Lyapunov exponents on a Stiefel manifold*, Physica D 156 (2001) 219–238. *[IV.9], [IV.10]*

Ch. Brouder, *Runge–Kutta methods and renormalization*, Euro. Phys. J. C 12 (2000) 521–534. *[III.1]*

Ch. Brouder, *Trees, Renormalization and Differential Equations*, BIT 44 (2004) 425–438. *[III.1]*

C.J. Budd & M.D. Piggott, *Geometric integration and its applications*, Handbook of Numerical Analysis XI (2003) 35–139. *[VIII.2]*

O. Buneman, *Time-reversible difference procedures*, J. Comput. Physics 1 (1967) 517–535. *[V.1]*

C. Burnton & R. Scherer, *Gauss–Runge–Kutta–Nyström methods*, BIT 38 (1998) 12–21. *[VI.10]*

K. Burrage & J.C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM J. Numer. Anal. 16 (1979) 46–57. *[VI.4]*

J.C. Butcher, *Coefficients for the study of Runge–Kutta integration processes*, J. Austral. Math. Soc. 3 (1963) 185–201. *[II.1]*

J.C. Butcher, *Implicit Runge–Kutta processes*, Math. Comput. 18 (1964a) 50–64. *[II.1]*

J.C. Butcher, *Integration processes based on Radau quadrature formulas*, Math. Comput. 18 (1964b) 233–244. *[II.1]*

J.C. Butcher, *The effective order of Runge–Kutta methods*, in J.Ll. Morris, ed., Proceedings of Conference on the Numerical Solution of Differential Equations, Lecture Notes in Math. 109 (1969) 133–139. *[V.3]*

J.C. Butcher, *An algebraic theory of integration methods*, Math. Comput. 26 (1972) 79–106. *[III.1], [III.3]*

J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations. Runge–Kutta and General Linear Methods*, John Wiley & Sons, Chichester, 1987. *[III.0], [III.1], [VI.7], [XV.8]*

J.C. Butcher, *Order and effective order*, Appl. Numer. Math. 28 (1998) 179–191. *[V.3]*

J.C. Butcher & J.M. Sanz-Serna, *The number of conditions for a Runge–Kutta method to have effective order p*, Appl. Numer. Math. 22 (1996) 103–111. *[III.1], [V.3]*

J.C. Butcher & G. Wanner, *Runge–Kutta methods: some historical notes*, Appl. Numer. Math. 22 (1996) 113–151. *[III.1]*

M.P. Calvo, *High order starting iterates for implicit Runge–Kutta methods: an improvement for variable-step symplectic integrators*, IMA J. Numer. Anal. 22 (2002) 153–166. *[VIII.6]*

M.P. Calvo & E. Hairer, *Accurate long-term integration of dynamical systems*, Appl. Numer. Math. 18 (1995a) 95–105. *[X.3]*

M.P. Calvo & E. Hairer, *Further reduction in the number of independent order conditions for symplectic, explicit Partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, Appl. Numer. Math. 18 (1995b) 107–114. *[III.3]*

M.P. Calvo, A. Iserles & A. Zanna, *Numerical solution of isospectral flows*, Math. Comput. 66 (1997) 1461–1486. *[IV.3]*

M.P. Calvo, A. Iserles & A. Zanna, *Conservative methods for the Toda lattice equations*, IMA J. Numer. Anal. 19 (1999) 509–523. *[IV.3]*

M.P. Calvo, M.A. López-Marcos & J.M. Sanz-Serna, *Variable step implementation of geometric integrators*, Appl. Numer. Math. 28 (1998) 1–6. *[VIII.2]*

M.P. Calvo, A. Murua & J.M. Sanz-Serna, *Modified equations for ODEs*, Contemporary Mathematics 172 (1994) 63–74. *[IX.9]*

M.P. Calvo & J.M. Sanz-Serna, *Variable steps for symplectic integrators*, In: Numerical Analysis 1991 (Dundee, 1991), 34–48, Pitman Res. Notes Math. Ser. 260, 1992. *[VIII.1]*

M.P. Calvo & J.M. Sanz-Serna, *The development of variable-step symplectic integrators, with application to the two-body problem*, SIAM J. Sci. Comput. 14 (1993) 936–952. *[V.3], [X.3]*

M.P. Calvo & J.M. Sanz-Serna, *Canonical B-series*, Numer. Math. 67 (1994) 161–175. *[VI.7]*

J. Candy & W. Rozmus, *A symplectic integration algorithm for separable Hamiltonian functions*, J. Comput. Phys. 92 (1991) 230–256. *[II.5]*

B. Cano & A. Durán, *Analysis of variable-stepsize linear multistep methods with special emphasis on symmetric ones*, Math. Comp. 72 (2003) 1769–1801. *[XV.7]*

B. Cano & A. Durán, *A technique to construct symmetric variable-stepsize linear multistep methods for second-order systems*, Math. Comp. 72 (2003) 1803–1816. *[XV.7]*

B. Cano & J.M. Sanz-Serna, *Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems*, IMA J. Numer. Anal. 18 (1998) 57–75. *[XV.5]*

R. Car & M. Parrinello, *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett. 55 (1985) 2471–2474. *[IV.9]*

J.R. Cash, *A class of implicit Runge–Kutta methods for the numerical integration of stiff ordinary differential equations*, J. Assoc. Comput. Mach. 22 (1975) 504–511. *[II.3]*

A. Cayley, *On the theory of the analytic forms called trees*, Phil. Magazine XIII (1857) 172–176. *[III.6]*

E. Celledoni & A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting*, IMA J. Numer. Anal. 21 (2001) 463–488. *[IV.8]*

R.P.K. Chan, *On symmetric Runge–Kutta methods of high order*, Computing 45 (1990) 301–309. *[VI.10]*

P.J. Channell & J.C. Scovel, *Integrators for Lie–Poisson dynamical systems*, Phys. D 50 (1991) 80–88. *[VII.5]*

P.J. Channell & J.C. Scovel, *Symplectic integration of Hamiltonian systems*, Nonlinearity 3 (1990) 231–259. *[VI.5]*

S. Chaplygin, *A new case of motion of a heavy rigid body supported in one point* (Russian), Moscov Phys. Sect. 10, vol. 2 (1901). *[X.1]*

P. Chartier, E. Faou & A. Murua, *An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants*, Preprint, February 2005. *[VI.7], [VI.8], [IX.9]*

M.T. Chu, *Matrix differential equations: a continuous realization process for linear algebra problems*, Nonlinear Anal. 18 (1992) 1125–1146. *[IV.3]*

S. Cirilli, E. Hairer & B. Leimkuhler, *Asymptotic error analysis of the adaptive Verlet method*, BIT 39 (1999) 25–33. *[VIII.3]*

A. Clebsch, *Ueber die simultane Integration linearer partieller Differentialgleichungen*, Crelle Journal f.d. reine u. angew. Math. 65 (1866) 257–268. *[VII.3]*

D. Cohen, *Analysis and numerical treatment of highly oscillatory differential equations*, Doctoral Thesis, Univ. Geneva, 2004. *[XIII.10]*

D. Cohen, *Conservation properties of numerical integrators for highly oscillatory Hamiltonian systems*, Report, 2005. To appear in IMA J. Numer. Anal. *[XIII.10]*

D. Cohen, E. Hairer & Ch. Lubich, *Modulated Fourier expansions of highly oscillatory differential equations*, Found. Comput. Math. 3 (2003) 327–345. *[XIII.6]*

D. Cohen, E. Hairer & Ch. Lubich, *Numerical energy conservation for multi-frequency oscillatory differential equations*, Report, 2004. To appear in BIT. *[XIII.9]*

G.J. Cooper, *Stability of Runge–Kutta methods for trajectory problems*, IMA J. Numer. Anal. 7 (1987) 1–13. *[IV.2]*

J.G. van der Corput, *Zur Methode der stationären Phase, I. Einfache Integrale*, Compos. Math. 1 (1934) 15–38. *[XIV.4]*

M. Creutz & A. Gocksch, *Higher-order hybrid Monte Carlo algorithms*, Phys. Rev. Lett. 63 (1989) 9–12. *[II.4]*

P.E. Crouch & R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci. 3 (1993) 1–33. *[IV.8]*
M. Crouzeix, *Sur la B-stabilité des méthodes de Runge–Kutta*, Numer. Math. 32 (1979) 75–82. *[VI.4]*
M. Crouzeix & J. Rappaz, *On Numerical Approximation in Bifurcation Theory*, Masson, Paris, 1989. *[XIV.3]*
G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. 4 (1956) 33-53. *[XV.1]*
G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. of the Royal Inst. of Techn. Stockholm, Sweden, Nr. 130 (1959) 87 pp. *[XV.5], [XV.9]*
G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Numerical Analysis, Dundee 1975, Lecture Notes in Math. 506 (1975) 60–74. *[VI.8], [XV.4]*
G. Darboux, *Sur le problème de Pfaff*, extraît Bulletin des Sciences math. et astron. 2e série, vol. VI (1882); Gauthier-Villars, Paris, 1882. *[VII.3]*
I. Degani & J. Schiff, *RCMS: Right correction Magnus series approach for integration of linear ordinary differential equations with highly oscillatory solution*, Report, Weizmann Inst. Science, Rehovot, 2003. *[XIV.1]*
P. Deift, *Integrable Hamiltonian systems*, in P. Deift (ed.) et al., Dynamical systems and probabilistic methods in partial differential equations. AMS Lect. Appl. Math. 31 (1996) 103–138. *[X.1]*
P. Deift, L.C. Li & C. Tomei, *Matrix factorizations and integrable systems*, Comm. Pure Appl. Math. 42 (1989) 443–521. *[IV.3]*
P. Deift, L.C. Li & C. Tomei, *Symplectic aspects of some eigenvalue algorithms*, in A.S. Fokas & V.E. Zakharov (eds.), Important Developments in Soliton Theory, Springer 1993. *[IV.3]*
P. Deift, T. Nanda & C. Tomei, *Ordinary differential equations and the symmetric eigenvalue problem*, SIAM J. Numer. Anal. 20 (1983) 1–22. *[IV.3]*
P. Deuflhard, *A study of extrapolation methods based on multistep schemes without parasitic solutions*, Z. angew. Math. Phys. 30 (1979) 177–189. *[XIII.1], [XIII.2]*
L. Dieci & T. Eirola, *On smooth decompositions of matrices*, SIAM J. Matrix Anal. Appl. 20 (1999) 800–819. *[IV.9]*
L. Dieci, R.D. Russell & E.S. van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal. 31 (1994) 261–281. *[IV.9]*
L. Dieci, R.D. Russell & E.S. van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal. 34 (1997) 402–423. *[IV.9], [IV.10]*
F. Diele, L. Lopez & R. Peluso, *The Cayley transform in the numerical solution of unitary differential systems*, Adv. Comput. Math. 8 (1998) 317–334. *[IV.8]*
F. Diele, L. Lopez & T. Politi, *One step semi-explicit methods based on the Cayley transform for solving isospectral flows*, J. Comput. Appl. Math. 89 (1998) 219–223. *[IV.3]*
P.A.M. Dirac, *Note on exchange phenomena in the Thomas atom*, Proc. Cambridge Phil. Soc. 26 (1930) 376–385. *[IV.9], [VII.6]*
P.A.M. Dirac, *Generalized Hamiltonian dynamics*, Can. J. Math. 2 (1950) 129–148. *[VII.7]*
V. Druskin & L. Knizhnerman, *Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl. 2 (1995) 205–217. *[XIII.1]*
A. Dullweber, B. Leimkuhler & R. McLachlan, *Symplectic splitting methods for rigid body molecular dynamics*, J. Chem. Phys. 107, No. 15 (1997) 5840–5851. *[VII.4], [VII.5]*
W. E, *Analysis of the heterogeneous multiscale method for ordinary differential equations*, Comm. Math. Sci. 1 (2003) 423–436. *[VIII.4]*
A. Edelman, T.A. Arias & S.T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl. 20 (1998) 303–353. *[IV.9]*

B.L. Ehle, *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*, Research Report CSRR 2010 (1969), Dept. AACS, Univ. of Waterloo, Ontario, Canada. *[II.1]*

E. Eich-Soellner & C. Führer, *Numerical Methods in Multibody Dynamics*, B. G. Teubner Stuttgart, 1998. *[IV.4], [VII.1]*

T. Eirola, *Aspects of backward error analysis of numerical ODE's*, J. Comp. Appl. Math. 45 (1993), 65–73. *[IX.1]*

T. Eirola & O. Nevanlinna, *What do multistep methods approximate?*, Numer. Math. 53 (1988) 559–569. *[XV.2]*

T. Eirola & J.M. Sanz-Serna, *Conservation of integrals and symplectic structure in the integration of differential equations by multistep methods*, Numer. Math. 61 (1992) 281–290. *[XV.4]*

L.H. Eliasson, *Absolutely convergent series expansions for quasi periodic motions*, Math. Phys. Electron. J. 2, No.4, Paper 4, 33 p. (1996). *[X.2]*

K. Engø & S. Faltinsen, *Numerical integration of Lie–Poisson systems while preserving coadjoint orbits and energy*, SIAM J. Numer. Anal. 39 (2001) 128–145. *[VII.5]*

B. Engquist & Y. Tsai, *Heterogeneous multiscale methods for stiff ordinary differential equations*, Math. Comp. 74 (2005) 1707–1742. *[VIII.4]*

Ch. Engstler & Ch. Lubich, *Multirate extrapolation methods for differential equations with different time scales*, Computing 58 (1997) 173–185. *[VIII.4]*

L. Euler, *Recherches sur la connoissance mécanique des corps*, Histoire de l'Acad. Royale de Berlin, Année MDCCLVIII, Tom. XIV, p. 131–153. Opera Omnia Ser. 2, Vol. 8, p. 178–199. *[VII.5]*

L. Euler, *Du mouvement de rotation des corps solides autour d'un axe variable*, Hist. de l'Acad. Royale de Berlin, Tom. 14, Année MDCCLVIII, 154–193. Opera Omnia Ser. II, Vol. 8, 200–235. *[IV.1]*

L. Euler, *Problème : un corps étant attiré en raison réciproque carrée des distances vers deux points fixes donnés, trouver les cas où la courbe décrite par ce corps sera algébrique*, Mémoires de l'Académie de Berlin for 1760, pub. 1767, 228–249. *[X.1]*

L. Euler, *Theoria motus corporum solidorum seu rigidorum*, Rostochii et Gryphiswaldiae A.F. Röse, MDCCLXV. Opera Omnia Ser. 2, Vol. 3-4. *[VII.5]*

L. Euler, *Institutionum Calculi Integralis*, Volumen Primum, *Opera Omnia,* Vol.XI. *[I.1]*

E. Faou, E. Hairer & T.-L. Pham, *Energy conservation with non-symplectic methods: examples and counter-examples*, submitted for publication. *[IX.9]*

E. Faou & Ch. Lubich, *A Poisson integrator for Gaussian wavepacket dynamics*, Report, 2004. To appear in Comp. Vis. Sci. *[VII.4], [VII.6]*

F. Fassò, *Comparison of splitting algorithms for the rigid body*, J. Comput. Phys. 189 (2003) 527–538. *[VII.5]*

K. Feng, *On difference schemes and symplectic geometry*, Proceedings of the 5-th Intern. Symposium on differential geometry & differential equations, August 1984, Beijing (1985) 42–58. *[VI.3]*

K. Feng, *Difference schemes for Hamiltonian formalism and symplectic geometry*, J. Comp. Math. 4 (1986) 279–289. *[VI.5]*

K. Feng, *Formal power series and numerical algorithms for dynamical systems*. In Proceedings of international conference on scientific computation, Hangzhou, China, Eds. Tony Chan & Zhong-Ci Shi, Series on Appl. Math. 1 (1991) 28–35. *[IX.1]*

K. Feng, *Collected Works (II)*, National Defense Industry Press, Beijing, 1995. *[XV.2]*

K. Feng & Z. Shang, *Volume-preserving algorithms for source-free dynamical systems*, Numer. Math. 71 (1995) 451–463. *[IV.3]*

K. Feng, H.M. Wu, M.-Z. Qin & D.L. Wang, *Construction of canonical difference schemes for Hamiltonian formalism via generating functions*, J. Comp. Math. 7 (1989) 71–96. *[VI.5]*

E. Fermi, J. Pasta & S. Ulam, *Studies of nonlinear problems*, Los Alamos Report No. LA-1940 (1955), later published in E. Fermi: Collected Papers (Chicago 1965), and Lect. Appl. Math. 15, 143 (1974). *[I.5]*

B. Fiedler & J. Scheurle, *Discretization of homoclinic orbits, rapid forcing and "invisible" chaos*, Mem. Amer. Math. Soc. 119, no. 570, 1996. *[IX.1]*

C.M. Field & F.W. Nijhoff, *A note on modified Hamiltonians for numerical integrations admitting an exact invariant*, Nonlinearity 16 (2003) 1673–1683. *[IX.11]*

L.N.G. Filon, *On a quadrature formula for trigonometric integrals*, Proc. Royal Soc. Edinburgh 49 (1928) 38–47. *[XIV.1]*

H. Flaschka, *The Toda lattice. II. Existence of integrals*, Phys. Rev. B 9 (1974) 1924–1925. *[IV.3]*

J. Ford, *The Fermi–Pasta–Ulam problem: paradox turns discovery*, Physics Reports 213 (1992) 271–310. *[I.5]*

E. Forest, *Canonical integrators as tracking codes*, AIP Conference Proceedings 184 (1989) 1106–1136. *[II.4]*

E. Forest, *Sixth-order Lie group integrators*, J. Comput. Physics 99 (1992) 209–213. *[V.3]*

E. Forest & R.D. Ruth, *Fourth-order symplectic integration*, Phys. D 43 (1990) 105–117. *[II.5]*

J. Frenkel, *Wave Mechanics, Advanced General Theory*, Clarendon Press, Oxford, 1934. *[IV.9], [VII.6]*

L. Galgani, A. Giorgilli, A. Martinoli & S. Vanzini, *On the problem of energy equipartition for large systems of the Fermi–Pasta–Ulam type: analytical and numerical estimates*, Physica D 59 (1992), 334–348. *[I.5]*

M.J. Gander, *A non spiraling integrator for the Lotka Volterra equation*, Il Volterriano 4 (1994) 21–28. *[VII.7]*

B. García-Archilla, J.M. Sanz-Serna & R.D. Skeel, *Long-time-step methods for oscillatory differential equations*, SIAM J. Sci. Comput. 20 (1999) 930–963. *[VIII.4], [XIII.1], [XIII.2], [XIII.4]*

L.M. Garrido, *Generalized adiabatic invariance*, J. Math. Phys. 5 (1964) 355–362. *[XIV.1]*

W. Gautschi, *Numerical integration of ordinary differential equations based on trigonometric polynomials*, Numer. Math. 3 (1961) 381–397. *[XIII.1]*

Z. Ge & J.E. Marsden, *Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators*, Phys. Lett. A 133 (1988) 134–139. *[VII.5], [IX.9]*

C.W. Gear & D.R. Wells, *Multirate linear multistep methods*, BIT 24 (1984) 484–502. *[VIII.4]*

W. Gentzsch & A. Schlüter, *Über ein Einschrittverfahren mit zyklischer Schrittweitenänderung zur Lösung parabolischer Differentialgleichungen*, ZAMM 58 (1978), T415–T416. *[II.4]*

S. Gill, *A process for the step-by-step integration of differential equations in an automatic digital computing machine*, Proc. Cambridge Philos. Soc. 47 (1951) 95–108. *[III.1], [VIII.5]*

A. Giorgilli & U. Locatelli, *Kolmogorov theorem and classical perturbation theory*, Z. Angew. Math. Phys. 48 (1997) 220–261. *[X.2]*

B. Gladman, M. Duncan & J. Candy, *Symplectic integrators for long-term integrations in celestial mechanics*, Celestial Mechanics and Dynamical Astronomy 52 (1991) 221–240. *[VIII.1]*

D. Goldman & T.J. Kaper, *Nth-order operator splitting schemes and nonreversible systems*, SIAM J. Numer. Anal. 33 (1996) 349–367. *[III.3]*

G.H. Golub & C.F. Van Loan, *Matrix Computations, 2nd edition*, John Hopkins Univ. Press, Baltimore and London, 1989. *[IV.4]*

O. Gonzalez, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Sci. 6 (1996) 449–467. *[V.5]*

O. Gonzalez, D.J. Higham & A.M. Stuart, *Qualitative properties of modified equations*. IMA J. Numer. Anal. 19 (1999) 169–190. *[IX.5]*

O. Gonzalez & J.C. Simo, *On the stability of symplectic and energy-momentum algorithms for nonlinear Hamiltonian systems with symmetry*, Comput. Methods Appl. Mech. Eng. 134 (1996) 197–222. *[V.5]*

D.N. Goryachev, *On the motion of a heavy rigid body with an immobile point of support in the case $A = B = 4C$* (Russian), Moscov Math. Collect. 21 (1899) 431–438. *[X.1]*

W.B. Gragg, *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations*, Thesis, Univ. of California; see also SIAM J. Numer. Anal. 2 (1965) 384–403. *[V.1]*

D.F. Griffiths & J.M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Stat. Comput. 7 (1986) 994–1008. *[IX.1]*

V. Grimm & M. Hochbruck, *Error analysis of exponential integrators for oscillatory second-order differential equations*, Preprint, 2005. *[XIII.4]*

W. Gröbner, *Die Liereihen und ihre Anwendungen*, VEB Deutscher Verlag der Wiss., Berlin 1960, 2nd ed. 1967. *[III.5]*

H. Grubmüller, H. Heller, A. Windemuth & K. Schulten, *Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions*, Mol. Sim. 6 (1991) 121–142. *[VIII.4], [XIII.1]*

A. Guillou & J.L. Soulé, *La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation*. Rev. Française Informat. Recherche Opŕationnelle 3 (1969) Ser. R-3, 17–44. *[II.1]*

M. Günther & P. Rentrop, *Multirate ROW methods and latency of electric circuits*. Appl. Numer. Math. 13 (1993) 83–102. *[VIII.4]*

F. Gustavson, *On constructing formal integrals of a Hamiltonian system near an equilibrium point*, Astron. J. 71 (1966) 670–686. *[I.3]*

J. Hadamard, *Sur l'itération et les solutions asymptotiques des équations différentielles*, Bull. Soc. Math. France 29 (1901) 224–228. *[XII.3]*

W.W. Hager, *Runge–Kutta methods in optimal control and the transformed adjoint system*, Numer. Math. 87 (2000) 247–282. *[VI.10]*

E. Hairer, *Backward analysis of numerical integrators and symplectic methods*, Annals of Numerical Mathematics 1 (1994) 107–132. *[VI.7]*

E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math. 25 (1997) 219–227. *[VIII.2]*

E. Hairer, *Backward error analysis for multistep methods*, Numer. Math. 84 (1999) 199–232. *[IX.9], [XV.3]*

E. Hairer, *Symmetric projection methods for differential equations on manifolds*, BIT 40 (2000) 726–734. *[V.4]*

E. Hairer, *Geometric integration of ordinary differential equations on manifolds*, BIT 41 (2001) 996–1007. *[V.4]*

E. Hairer, *Global modified Hamiltonian for constrained symplectic integrators*, Numer. Math. 95 (2003) 325–336. *[IX.5]*

E. Hairer & M. Hairer, *GniCodes – Matlab programs for geometric numerical integration*, In: Frontiers in numerical analysis (Durham, 2002), Springer Berlin, Universitext (2003), 199–240. *[VIII.6]*

E. Hairer & P. Leone, *Order barriers for symplectic multi-value methods*. In: Numerical analysis 1997, Proc. of the 17th Dundee Biennial Conference, June 24-27, 1997, D.F. Griffiths, D.J. Higham & G.A. Watson (eds.), Pitman Research Notes in Mathematics Series 380 (1998), 133–149. *[XV.4], [XV.8]*

E. Hairer & P. Leone, *Some properties of symplectic Runge–Kutta methods*, New Zealand J. of Math. 29 (2000) 169–175. *[IV.2]*

E. Hairer & Ch. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math. 76 (1997), pp. 441–462. Erratum: http://www.unige.ch/math/folks/hairer/ *[IX.7], [X.5]*

E. Hairer & Ch. Lubich, *Invariant tori of dissipatively perturbed Hamiltonian systems under symplectic discretization*, Appl. Numer. Math. 29 (1999) 57–71. *[XII.1], [XII.5]*

E. Hairer & Ch. Lubich, *Asymptotic expansions and backward analysis for numerical integrators*, Dynamics of Algorithms (Minneapolis, MN, 1997), IMA Vol. Math. Appl. 118, Springer, New York (2000) 91–106. *[IX.1]*

E. Hairer & Ch. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal. 38 (2000a) 414-441. *[XIII.1], [XIII.2], [XIII.5], [XIII.7]*

E. Hairer & Ch. Lubich, *Energy conservation by Störmer-type numerical integrators*, in: G.F. Griffiths, G.A. Watson (eds.), Numerical Analysis 1999, CRC Press LLC (2000b) 169–190. *[XIII.8]*

E. Hairer & Ch. Lubich, *Symmetric multistep methods over long times*, Numer. Math. 97 (2004) 699–723. *[XV.3], [XV.5], [XV.6]*

E. Hairer, Ch. Lubich & M. Roche, *The numerical solution of differential-algebraic systems by Runge–Kutta methods*, Lecture Notes in Math. 1409, Springer-Verlag, 1989. *[VII.1]*

E. Hairer, Ch. Lubich & G. Wanner, *Geometric numerical integration illustrated by the Störmer–Verlet method*, Acta Numerica (2003) 399–450. *[I.1]*

E. Hairer, S.P. Nørsett & G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd edition*, Springer Series in Computational Mathematics **8**, Springer Berlin, 1993. *[II.1]*

E. Hairer & G. Söderlind, *Explicit, time reversible, adaptive step size control*, Submitted for publication, 2004. *[VIII.3], [IX.6]*

E. Hairer & D. Stoffer, *Reversible long-term integration with variable stepsizes*, SIAM J. Sci. Comput. 18 (1997) 257–269. *[VIII.3]*

E. Hairer & G. Wanner, *On the Butcher group and general multi-value methods*, Computing 13 (1974) 1–15. *[III.1]*

E. Hairer & G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, 2nd edition*, Springer Series in Computational Mathematics **14**, Springer-Verlag Berlin, 1996. *[II.1], [III.0], [IV.2], [IV.4], [IV.5], [IV.9], [IV.10], [VI.4], [VI.10], [VII.1], [VIII.6], [IX.5], [XIII.2], [XV.4], [XV.9]*

E. Hairer & G. Wanner, *Analysis by Its History, 2nd printing*, Undergraduate Texts in Mathematics, Springer-Verlag New York, 1997. *[IX.7]*

M. Hall, jr., *A basis for free Lie rings and higher commutators in free groups*, Proc. Amer. Math. Soc. 1 (1950) 575–581. *[III.3]*

Sir W.R. Hamilton, *On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function*, Phil. Trans. Roy. Soc. Part II for 1834, 247–308; Math. Papers, Vol. II, 103–161. *[VI.1], [VI.5]*

P.C. Hammer & J.W. Hollingsworth, *Trapezoidal methods of approximating solutions of differential equations*, MTAC 9 (1955) 92–96. *[II.1]*

E.J. Haug, *Computer Aided Kinematics and Dynamics of Mechanical Systems, Volume I: Basic Methods*, Allyn & Bacon, Boston, 1989. *[VII.5]*

F. Hausdorff, *Die symbolische Exponentialformel in der Gruppentheorie*, Berichte der Sächsischen Akad. der Wissensch. 58 (1906) 19–48. *[III.4]*

A. Hayli, *Le problème des N corps dans un champ extérieur application a l'évolution dynamique des amas ouverts - I*, Bulletin Astronomique 2 (1967) 67–89. *[VIII.4]*

R.B. Hayward, *On a Direct Method of estimating Velocities, Accelerations, and all similar Quantities with respect to Axes moveable in any Space, with Applications*, Cambridge Phil. Trans. vol. X (read 1856, publ. 1864) 1–20. *[VII.5]*

E.J. Heller, *Time dependent approach to semiclassical dynamics*, J. Chem. Phys. 62 (1975) 1544–1555. *[VII.6]*

E.J. Heller, *Time dependent variational approach to semiclassical dynamics*, J. Chem. Phys. 64 (1976) 63–73. *[VII.6]*

M. Hénon & C. Heiles, *The applicability of the third integral of motion : some numerical experiments*, Astron. J. 69 (1964) 73–79. *[I.3]*

J. Henrard, *The adiabatic invariant in classical mechanics*, Dynamics reported, New series. Vol. 2, Springer, Berlin (1993) 117–235. *[XIV.1]*

P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York 1962. *[VIII.5]*

J. Hersch, *Contribution à la méthode aux différences*, Z. angew. Math. Phys. 9a (1958) 129–180. *[XIII.1]*

K. Heun, *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Zeitschr. für Math. u. Phys. 45 (1900) 23–38. *[II.1]*

D.J. Higham, *Time-stepping and preserving orthogonality*, BIT 37 (1997) 24–36. *[IV.9]*

N.J. Higham, *The accuracy of floating point summation*, SIAM J. Sci. Comput. 14 (1993) 783–799. *[VIII.5]*

M. Hochbruck & Ch. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal. 34 (1997) 1911–1925. *[XIII.1]*

M. Hochbruck & Ch. Lubich, *A Gautschi-type method for oscillatory second-order differential equations*, Numer. Math. 83 (1999a) 403–426. *[VIII.4], [XIII.1], [XIII.2], [XIII.4]*

M. Hochbruck & Ch. Lubich, *Exponential integrators for quantum-classical molecular dynamics*, BIT 39 (1999b) 620–645. *[VIII.4], [XIV.1], [XIV.4]*

T. Holder, B. Leimkuhler & S. Reich, *Explicit variable step-size and time-reversible integration*, Appl. Numer. Math. 39 (2001) 367–377. *[VIII.3]*

H. Hopf, *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. 42 (1941) 22–52. *[III.1]*

W. Huang & B. Leimkuhler, *The adaptive Verlet method*, SIAM J. Sci. Comput. 18 (1997) 239–256. *[VIII.2], [VIII.3]*

P. Hut, J. Makino & S. McMillan, *Building a better leapfrog*, Astrophys. J. 443 (1995) L93–L96. *[VIII.3]*

K.J. In't Hout, *A new interpolation procedure for adapting Runge–Kutta methods to delay differential equations*, BIT 32 (1992) 634–649. *[VIII.6]*

A. Iserles, *Solving linear ordinary differential equations by exponentials of iterated commutators*, Numer. Math. 45 (1984) 183–199. *[II.4]*

A. Iserles, *On the global error of discretization methods for highly-oscilatory ordinary differential equations*, BIT 42 (2002) 561–599. *[XIV.1]*

A. Iserles, *On the method of Neumann series for highly oscillatory equations*, BIT 44 (2004) 473–488. *[XIV.1]*

A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett & A. Zanna, *Lie-group methods*, Acta Numerica (2000) 215–365. *[IV.8]*

A. Iserles & S.P. Nørsett, *On the solution of linear differential equations in Lie groups*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. 357 (1999) 983–1019. *[IV.7], [IV.10]*

A. Iserles & S.P. Nørsett, *On the numerical quadrature of highly-oscillating integrals I: Fourier transforms*, IMA J. Numer. Anal. 24 (2004) 365–391. *[XIV.1]*

T. Itoh & K. Abe, *Hamiltonian-conserving discrete canonical equations based on variational difference quotients*, J. Comput. Phys. 76 (1988) 85–102. *[V.5]*

J.A. Izaguirre, S. Reich & R.D. Skeel, *Longer time steps for molecular dynamics*, J. Chem. Phys. 110 (1999) 9853–9864. *[XIII.1], [XIV.4]*

C.G.J. Jacobi, *Über diejenigen Probleme der Mechanik, in welchen eine Kräftefunction existirt, und über die Theorie der Störungen*, manuscript from 1836 or 1837, published posthumely in *Werke*, vol. 5, 217–395. *[VI.2]*

C.G.J. Jacobi, *Über die Reduktion der Integration der partiellen Differentialgleichungen erster Ordnung zwischen irgend einer Zahl Variablen auf die Integration eines einzigen Systemes gewöhnlicher Differentialgleichungen*, Crelle Journal f.d. reine u. angew. Math. 17 (1837) 97–162; K. Weierstrass, ed., C.G.J. Jacobi's Gesammelte Werke, vol. 4, pp. 57–127. *[VI.5]*

C.G.J. Jacobi, *Lettre adressée à M. le Président de l'Académie des Sciences*, Liouville J. math. pures et appl. 5 (1840) 350–355; *Werke*, vol. 5, pp. 3–189. *[IV.1]*

C.G.J. Jacobi, *Vorlesungen über Dynamik* (1842-43), Reimer, Berlin 1884. *[VI.1], [VI.5], [VI.6], [VI.10]*

C.G.J. Jacobi, *Nova methodus, aequationes differentiales partiales primi ordini inter numerum variabilium quemcunque propositas integrandi*, published posthumly in Crelle Journal f.d. reine u. angew. Math. 60 (1861) 1–181; *Werke*, vol. 5, pp. 3–189. *[III.5], [VII.2], [VII.3]*

T. Jahnke, *Numerische Verfahren für fast adiabatische Quantendynamik*, Doctoral Thesis, Univ. Tübingen, 2003. *[XIV.3]*

T. Jahnke, *Long-time-step integrators for almost-adiabatic quantum dynamics*, SIAM J. Sci. Comput. 25 (2004a) 2145–2164. *[XIV.1]*

T. Jahnke, *A long-time-step method for quantum-classical molecular dynamics*, Report, 2004b. *[XIV.3]*

T. Jahnke & Ch. Lubich, *Numerical integrators for quantum dynamics close to the adiabatic limit*, Numer. Math. 94 (2003), 289–314. *[XIV.1]*

L. Jay, *Collocation methods for differential-algebraic equations of index 3*, Numer. Math. 65 (1993) 407–421. *[VII.1]*

L. Jay, *Runge–Kutta type methods for index three differential-algebraic equations with applications to Hamiltonian systems*, Thesis No. 2658, 1994, Univ. Genève. *[VII.1]*

L. Jay, *Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems*, SIAM J. Numer. Anal. 33 (1996) 368–387. *[II.2], [VII.1]*

L. Jay, *Specialized Runge–Kutta methods for index 2 differential algebraic equations*, Math. Comp. (2005), to appear. *[IV.9]*

R. Jost, *Winkel- und Wirkungsvariable für allgemeine mechanische Systeme*, Helv. Phys. Acta 41 (1968) 965–968. *[X.1]*

A. Joye & C.-E. Pfister, *Superadiabatic evolution and adiabatic transition probability between two nondegenerate levels isolated in the spectrum*, J. Math. Phys. 34 (1993) 454–479. *[XIV.1]*

W. Kahan, *Further remarks on reducing truncation errors*, Comm. ACM 8 (1965) 40. *[VIII.5]*

W. Kahan & R.-C. Li, *Composition constants for raising the orders of unconventional schemes for ordinary differential equations*, Math. Comput. 66 (1997) 1089–1099. *[V.3], [V.6]*

B. Karasözen, *Poisson integrators*, Math. Comp. Modelling 40 (2004) 1225–1244. *[VII.4]*

T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, 1980. *[VII.6]*

J. Kepler, *Astronomia nova αιτιολογητός seu Physica celestis, traditia commentariis de motibus stellae Martis, ex observationibus G. V. Tychonis Brahe*, Prague 1609. *[I.2]*

H. Kinoshita, H. Yoshida & H. Nakai, *Symplectic integrators and their application to dynamical astronomy*, Celest. Mech. & Dynam. Astr. 50 (1991) 59–71. *[V.3]*

U. Kirchgraber, *Multi-step methods are essentially one-step methods*, Numer. Math. 48 (1986) 85–90. *[XV.2]*

U. Kirchgraber, F. Lasagni, K. Nipp & D. Stoffer, *On the application of invariant manifold theory, in particular to numerical analysis*, Internat. Ser. Numer. Math. 97, Birkhäuser, Basel, 1991, 189–197. *[XII.3]*

U. Kirchgraber & E. Stiefel, *Methoden der analytischen Störungsrechnung und ihre Anwendungen*, Teubner, Stuttgart, 1978. *[XII.4]*

F. Klein, *Elementarmathematik vom höheren Standpunkte aus. Teil I: Arithmetik, Algebra, Analysis*, ausgearbeitet von E. Hellinger, Teubner, Leipzig, 1908; Vierte Auflage, Die Grundlehren der mathematischen Wissenschaften, Band 14 Springer-Verlag, Berlin, 1933, reprinted 1968. *[VII.5]*

F. Klein & A. Sommerfeld, *Theorie des Kreisels*, Leipzig 1897. *[VII.5]*

O. Koch & Ch. Lubich, *Dynamical low rank approximation*, Preprint, 2005. *[IV.9]*

A.N. Kolmogorov, *On conservation of conditionally periodic motions under small perturbations of the Hamiltonian*, Dokl. Akad. Nauk SSSR 98 (1954) 527–530. *[X.2], [X.5]*

A.N. Kolmogorov, *General theory of dynamical systems and classical mechanics*, Proc. Int. Congr. Math. Amsterdam 1954, Vol. 1, 315–333. *[X.2], [X.5]*

P.-V. Koseleff, *Exhaustive search of symplectic integrators using computer algebra*, Integration algorithms and classical mechanics, Fields Inst. Commun. 10 (1996) 103–120. *[V.3]*

S. Kovalevskaya (Kowalevski), *Sur le problème de la rotation d'un corps solide autour d'un point fixe*, Acta Math. 12 (1889) 177–232. *[X.1]*

V.V. Kozlov, *Integrability and non-integrability in Hamiltonian mechanics*, Uspekhi Mat. Nauk 38 (1983) 3–67. *[X.1]*

D. Kreimer, *On the Hopf algebra structure of perturbative quantum field theory*, Adv. Theor. Math. Phys. 2 (1998) 303–334. *[III.1]*

N.M. Krylov & N.N. Bogoliubov, *Application des méthodes de la mécanique non linéaire à la théorie des oscillations stationnaires*, Edition de l'Académie des Sciences de la R.S.S. d'Ukraine, 1934. *[XII.4]*

W. Kutta, *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Zeitschr. für Math. u. Phys. 46 (1901) 435–453. *[II.1]*

R.A. LaBudde & D. Greenspan, *Discrete mechanics – a general treatment*, J. Comput. Phys. 15 (1974) 134–167. *[V.5]*

R.A. LaBudde & D. Greenspan, *Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion. Parts I and II*, Numer. Math. 25 (1976) 323–346 and 26 (1976) 1–26. *[V.5]*

M.P. Laburta, *Starting algorithms for IRK methods*, J. Comput. Appl. Math. 83 (1997) 269–288. *[VIII.6]*

M.P. Laburta, *Construction of starting algorithms for the RK-Gauss methods*, J. Comput. Appl. Math. 90 (1998) 239–261. *[VIII.6]*

J.-L. Lagrange, *Applications de la methode exposée dans le mémoire précédent a la solution de différents problèmes de dynamique*, 1760, Oeuvres Vol. 1, 365–468. *[VI.1], [VI.2]*

J.L. Lagrange, *Recherches sur le mouvement d'un corps qui est attiré vers deux centres fixes* (1766), Œuvres, tome II, Gauthier-Villars, Paris 1868, 67–124. *[X.1]*

J.-L. Lagrange, *Mécanique analytique*, Paris 1788. *[VI.1]*

J.D. Lambert & I.A. Watson, *Symmetric multistep methods for periodic initial value problems*, J. Inst. Maths. Applics. 18 (1976) 189–202. *[XV.1], [XV.9]*

C. Lanczos, *The Variational Principles of Mechanics*, University of Toronto Press, Toronto, 1949. (Fourth edition 1970). *[VI.6]*

P.S. Laplace, *Traité de mécanique céleste II*, 1799, see *Œuvres* I, p. 183. *[I.6]*

F.M. Lasagni, *Canonical Runge–Kutta methods*, ZAMP 39 (1988) 952–953. *[VI.4], [VI.5], [VI.7]*

J.D. Lawson, *Generalized Runge-Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal. 4 (1967) 372–380. *[XIV.1]*

P.D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Commun. Pure Appl. Math. 21 (1968) 467–490. *[IV.3]*

B. Leimkuhler & S. Reich, *Symplectic integration of constrained Hamiltonian systems*, Math. Comp. 63 (1994) 589–605. *[VII.1]*

B. Leimkuhler & S. Reich, *A reversible averaging integrator for multiple time-scale dynamics*, J. Comput. Phys. 171 (2001) 95–114. *[VIII.4]*

B. Leimkuhler & S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monographs on Applied and Computational Mathematics **14**, Cambridge University Press, Cambridge, 2004. *[VI.3]*

B.J. Leimkuhler & R.D. Skeel, *Symplectic numerical integrators in constrained Hamiltonian systems*, J. Comput. Phys. 112 (1994) 117–125. *[VII.1]*

A. Lenard, *Adiabatic invariance to all orders*, Ann. Phys. 6 (1959) 261–276. *[XIV.1]*

P. Leone, *Symplecticity and Symmetry of General Integration Methods*, Thèse, Section de Mathématiques, Université de Genève, 2000. *[VI.8]*

T. Levi-Civita, *Sur la résolution qualitative du problème restreint des trois corps*, Acta Math. 30 (1906) 305–327. *[VIII.2]*

T. Levi-Civita, *Sur la régularisation du problème des trois corps*, Acta Math. 42 (1920) 99–144. *[VIII.2]*

D. Lewis & J.C. Simo, *Conserving algorithms for the dynamics of Hamiltonian systems on Lie groups*, J. Nonlinear Sci. 4 (1994) 253–299. *[IV.8], [V.5]*

D. Lewis & J.C. Simo, *Conserving algorithms for the N-dimensional rigid body*, Fields Inst. Com. 10 (1996) 121–139. *[V.5]*

S. Lie, *Zur Theorie der Transformationsgruppen*, Christ. Forh. Aar. 1888, Nr. 13, 6 pages, Christiania 1888; Gesammelte Abh. vol. 5, p. 553–557. *[VII.2], [VII.3]*

J. Liouville, *Note à l'occasion du mémoire précédent (de M. E. Bour)*, J. Math. Pures et Appliquées 20 (1855) 201–202. *[X.1]*

L. Lopez & T. Politi, *Applications of the Cayley approach in the numerical solution of matrix differential systems on quadratic groups*, Appl. Numer. Math. 36 (2001) 35–55. *[IV.8]*

M.A. López-Marcos, J.M. Sanz-Serna & R.D. Skeel, *Cheap enhancement of symplectic integrators*, Numerical analysis 1995 (Dundee), Pitman Res. Notes Math. Ser. 344, Longman, Harlow, 1996, 107–122. *[V.3]*

K. Lorenz, T. Jahnke & Ch. Lubich, *Adiabatic integrators for highly oscillatory second order linear differential equations with time-varying eigendecomposition*, BIT 45 (2005) 91–115. *[XIV.1], [XIV.2]*

A.J. Lotka, *The Elements of Physical Biology*, Williams & Wilkins, Baltimore, 1925. Reprinted 1956 under the title *Elements of mathematical biology* by Dover, New York. *[I.1]*

Ch. Lubich, *Integration of stiff mechanical systems by Runge-Kutta methods*, Z. Angew. Math. Phys. 44 (1993) 1022–1053. *[XIV.3]*

Ch. Lubich, *On dynamics and bifurcations of nonlinear evolution equations under numerical discretization*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems (B. Fiedler, ed.), Springer, Berlin, 2001, 469–500. *[XII.3]*

Ch. Lubich, *A variational splitting integrator for quantum molecular dynamics*, Appl. Numer. Math. 48 (2004) 355–368. *[VII.4]*

Ch. Lubich, *On variational approximations in quantum molecular dynamics*, Math. Comp. 74 (2005) 765–779. *[VII.6]*

R. MacKay, *Some aspects of the dynamics of Hamiltonian systems*, in: D.S. Broomhead & A. Iserles, eds., *The Dynamics of Numerics and the Numerics of Dynamics*, Clarendon Press, Oxford, 1992, 137–193. *[VI.6]*

S. Maeda, *Canonical structure and symmetries for discrete systems*, Math. Japonica 25 (1980) 405–420. *[VI.6]*

S. Maeda, *Lagrangian formulation of discrete systems and concept of difference space*, Math. Japonica 27 (1982) 345–356. *[VI.6]*

W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math. VII (1954) 649–673. *[IV.7]*

G. Marchuk, *Some applications of splitting-up methods to the solution of mathematical physics problems*, Aplikace Matematiky 13 (1968) 103–132. *[II.5]*

J.E. Marsden, S. Pekarsky & S. Shkoller, *Discrete Euler-Poincaré and Lie-Poisson equations*, Nonlinearity 12 (1999) 1647–1662. *[VII.5]*

J.E. Marsden & T.S. Ratiu, *Introduction to Mechanics and Symmetry. A Basic Exposition of Classical Mechanical Systems*, Second edition, Texts in Applied Mathematics 17, Springer-Verlag, New York, 1999. *[IV.1]*

J.E. Marsden & M. West, *Discrete mechanics and variational integrators*, Acta Numerica 10 (2001) 1–158. *[VI.6]*

A.D. McLachlan, *A variational solution of the time-dependent Schrodinger equation*, Mol. Phys. 8 (1964) 39–44. *[VII.6]*

R.I. McLachlan, *Explicit Lie-Poisson integration and the Euler equations*, Phys. Rev. Lett. 71 (1993) 3043–3046. *[VII.4], [VII.5]*

R.I. McLachlan, *On the numerical integration of ordinary differential equations by symmetric composition methods*, SIAM J. Sci. Comput. 16 (1995) 151–168. *[II.4], [II.5], [III.3], [V.3], [V.6]*

R.I. McLachlan, *Composition methods in the presence of small parameters*, BIT 35 (1995b) 258–268. *[V.3]*

R.I. McLachlan, *More on symplectic integrators*, in *Integration Algorithms and Classical Mechanics* 10, J.E. Marsden, G.W. Patrick & W.F. Shadwick, eds., Amer. Math. Soc., Providence, R.I. (1996) 141–149. *[V.3]*

R.I. McLachlan, *Featured review of Geometric Numerical Integration by E. Hairer, C. Lubich, and G. Wanner*, SIAM Review 45 (2003) 817–821. *[VII.5]*

R.I. McLachlan & P. Atela, *The accuracy of symplectic integrators*, Nonlinearity 5 (1992) 541–562. *[V.3]*

R.I. McLachlan & G.R.W. Quispel, *Splitting methods*, Acta Numerica 11 (2002) 341–434. *[VII.4]*

R.I. McLachlan, G.R.W. Quispel & N. Robidoux, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond., Ser. A, 357 (1999) 1021–1045. *[V.5]*

R.I. McLachlan & C. Scovel, *Equivariant constrained symplectic integration*, J. Nonlinear Sci. 5 (1995) 233–256. *[VII.5]*

R.I. McLachlan & A. Zanna, *The discrete Moser–Veselov algorithm for the free rigid body, revisited*, Found. Comput. Math. 5 (2005) 87–123. *[VII.5], [IX.11]*

R.J.Y. McLeod & J.M. Sanz-Serna, *Geometrically derived difference formulae for the numerical integration of trajectory problems*, IMA J. Numer. Anal. 2 (1982) 357–370. *[VIII.2]*

V.L. Mehrmann, *The Autonomous Linear Quadratic Control Problem. Theory and Numerical Solution*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991. *[IV.9]*

R.H. Merson, *An operational method for the study of integration processes*, Proc. Symp. Data Processing, Weapons Research Establishment, Salisbury, Australia (1957) 110-1 to 110-25. *[III.1]*

A. Messiah, *Quantum Mechanics*, Dover Publ., 1999 (reprint of the two-volume edition published by Wiley, 1961-1962). *[VII.6]*

S. Miesbach & H.J. Pesch, *Symplectic phase flow approximation for the numerical integration of canonical systems*, Numer. Math. 61 (1992) 501–521. *[VI.5]*

P.C. Moan, *On rigorous modified equations for discretizations of ODEs*, Report, 2005. *[IX.7]*

O. Møller, *Quasi double-precision in floating point addition*, BIT 5 (1965) 37–50 and 251–255. *[VIII.5]*

A. Morbidelli & A. Giorgilli, *Superexponential stability of KAM Tori*, J. Stat. Phys. 78 (1995) 1607–1617. *[X.2]*

J. Moser, *Review* MR 20-4066, Math. Rev., 1959. *[X.5]*

J. Moser, *On invariant curves of area-preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen, II. Math.-Phys. Kl. 1962, 1–20. *[X.5]*

J. Moser, *Lectures on Hamiltonian systems*, Mem. Am. Math. Soc. 81 (1968) 1–60. *[IX.3]*

J. Moser, *Stable and Random Motions in Dynamical Systems*, Annals of Mathematics Studies. No. 77. Princeton University Press, 1973. *[XI.2]*

J. Moser, *Finitely many mass points on the line under the influence of an exponential potential — an integrable system*, Dyn. Syst., Theor. Appl., Battelle Seattle 1974 Renc., Lect. Notes Phys. 38 (1975) 467–497. *[X.1]*

J. Moser, *Is the solar system stable?*, Mathematical Intelligencer 1 (1978) 65–71. *[X.0]*

J. Moser & A.P. Veselov, *Discrete versions of some classical integrable systems and factorization of matrix polynomials*, Commun. Math. Phys. 139 (1991) 217–243. *[VII.5]*

H. Munthe-Kaas, *Lie Butcher theory for Runge–Kutta methods*, BIT 35 (1995) 572–587. *[IV.8]*

H. Munthe-Kaas, *Runge–Kutta methods on Lie groups*, BIT 38 (1998) 92–111. *[IV.8]*

H. Munthe-Kaas, *High order Runge–Kutta methods on manifolds*, J. Appl. Num. Maths. 29 (1999) 115–127. *[IV.8]*

H. Munthe-Kaas & B. Owren, *Computations in a free Lie algebra*, Phil. Trans. Royal Soc. A 357 (1999) 957–981. *[IV.7]*

A. Murua, *Métodos simplécticos desarrollables en P-series*, Doctoral Thesis, Univ. Valladolid, 1994. *[IX.3]*

A. Murua, *On order conditions for partitioned symplectic methods*, SIAM J. Numer. Anal. 34 (1997) 2204–2211. *[IX.11]*

A. Murua, *Formal series and numerical integrators, Part I: Systems of ODEs and symplectic integrators*, Appl. Numer. Math. 29 (1999) 221–251. *[IX.11]*

A. Murua & J.M. Sanz-Serna, *Order conditions for numerical integrators obtained by composing simpler integrators*, Philos. Trans. Royal Soc. London, ser. A 357 (1999) 1079–1100. *[III.1], [III.3], [V.3]*

A.I. Neishtadt, *The separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech. 48 (1984) 133–139. *[XIV.2]*

N.N. Nekhoroshev, *An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems*, Russ. Math. Surveys 32 (1977) 1–65. *[X.2], [X.4]*

N.N. Nekhoroshev, *An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems. II.* (Russian), Tr. Semin. Im. I.G. Petrovskogo 5 (1979) 5–50. *[X.4]*

G. Nenciu, *Linear adiabatic theory. Exponential estimates*, Commun. Math. Phys. 152 (1993) 479–496. *[XIV.1]*

P. Nettesheim & S. Reich, *Symplectic multiple-time-stepping integrators for quantum-classical molecular dynamics*, in P. Deuflhard et al. (eds.), Computational Molecular Dynamics: Challenges, Methods, Ideas, Springer, Berlin 1999, 412–420. *[VIII.4]*

I. Newton, *Philosophiae Naturalis Principia Mathematica*, Londini anno MDCLXXXVII, 1687. *[I.2], [VI.1], [X.1]*

I. Newton, *Second edition of the Principia*, 1713. *[I.2], [X.1]*

K. Nipp & D. Stoffer, *Attractive invariant manifolds for maps: existence, smoothness and continuous dependence on the map*, Research Report No. 92–11, SAM, ETH Zürich, 1992. *[XII.3]*

K. Nipp & D. Stoffer, *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type. I: RK-methods*, Numer. Math. 70 (1995) 245–257. *[XII.3]*

K. Nipp & D. Stoffer, *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type. II: Linear multistep methods*, Numer. Math. 74 (1996) 305–323. *[XII.3]*

E. Noether, *Invariante Variationsprobleme*, Nachr. Akad. Wiss. Göttingen, Math.-Phys. Kl. (1918) 235–257. *[VI.6]*

E.J. Nyström, *Ueber die numerische Integration von Differentialgleichungen*, Acta Soc. Sci. Fenn. 50 (1925) 1–54. *[II.2]*

E. Oja, *Neural networks, principal components, and subspaces*, Int. J. Neural Syst. 1 (1989) 61–68. *[IV.9]*

D. Okunbor & R.D. Skeel, *Explicit canonical methods for Hamiltonian systems*, Math. Comp. 59 (1992) 439–455. *[VI.4]*

D.I. Okunbor & R.D. Skeel, *Canonical Runge–Kutta–Nyström methods of orders five and six*, J. Comp. Appl. Math. 51 (1994) 375–382. *[V.3]*

F.W.J. Olver, *Asymptotics and Special Functions*, Academic Press, 1974. *[XIV.4]*

P.J. Olver, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Mathematics **107**, Springer-Verlag, New York, 1986. *[IV.6]*

B. Owren & A. Marthinsen, *Runge–Kutta methods adapted to manifolds and based on rigid frames*, BIT 39 (1999) 116–142. *[IV.8]*

B. Owren & A. Marthinsen, *Integration methods based on canonical coordinates of the second kind*, Numer. Math. 87 (2001) 763–790. *[IV.8]*

A.M. Perelomov, *Selected topics on classical integrable systems*, Troisième cycle de la physique, expanded version of lectures delivered in May 1995. *[VII.2]*

O. Perron, *Über Stabilität und asymptotisches Verhalten der Lösungen eines Systems endlicher Differenzengleichungen*, J. Reine Angew. Math. 161 (1929) 41–64. *[XII.3]*

A.D. Perry & S. Wiggins, *KAM tori are very sticky: Rigorous lower bounds on the time to move away from an invariant Lagrangian torus with linear flow*, Physica D 71 (1994) 102–121. *[X.2]*

H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste, Tome I*, Gauthier-Villars, Paris, 1892. *[VI.1], [X.1], [X.2]*

H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste, Tome II*, Gauthier-Villars, Paris, 1893. *[VI.1], [X.2]*

H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste. Tome III*, Gauthiers-Villars, Paris, 1899. *[VI.1], [VI.2]*

L. Poinsot, *Théorie nouvelle de la rotation des corps*, Paris 1834. *[VII.5]*

S.D. Poisson, *Sur la variation des constantes arbitraires dans les questions de mécanique*, J. de l'Ecole Polytechnique vol. 8, 15e cahier (1809) 266–344. *[VII.2]*

B. van der Pol, *Forced oscillations in a system with non-linear resistance*, Phil. Mag. 3, (1927), 65–80; *Papers* vol. I, 361–376. *[XII.4]*

J. Pöschel, *Nekhoroshev estimates for quasi-convex Hamiltonian systems*, Math. Z. 213 (1993) 187–216. *[X.2]*

F.A. Potra & W.C. Rheinboldt, *On the numerical solution of Euler–Lagrange equations*, Mech. Struct. & Mech. 19 (1991) 1–18. *[IV.5]*

M.-Z. Qin & W.-J. Zhu, *Volume-preserving schemes and numerical experiments*, Comput. Math. Appl. 26 (1993) 33–42. *[VI.9]*

G.D. Quinlan, *Resonances and instabilities in symmetric multistep methods*, Report, 1999, available on http://xxx.lanl.gov/abs/astro-ph/9901136 *[XV.7]*

G.D. Quinlan & S. Tremaine, *Symmetric multistep methods for the numerical integration of planetary orbits*, Astron. J. 100 (1990) 1694–1700. *[XV.1], [XV.7]*

G.R.W. Quispel, *Volume-preserving integrators*, Phys. Lett. A 206 (1995) 26–30. *[VI.9]*

S. Reich, *Symplectic integration of constrained Hamiltonian systems by Runge–Kutta methods*, Techn. Report 93-13 (1993), Dept. Comput. Sci., Univ. of British Columbia. *[VII.1]*

S. Reich, *Numerical integration of the generalized Euler equations*, Techn. Report 93-20 (1993), Dept. Comput. Sci., Univ. of British Columbia. *[VII.4]*

S. Reich, *Momentum conserving symplectic integrators*, Phys. D 76 (1994) 375–383. *[VII.5]*

S. Reich, *Symplectic integration of constrained Hamiltonian systems by composition methods*, SIAM J. Numer. Anal. 33 (1996a) 475–491. *[VII.1], [IX.5]*

S. Reich, *Enhancing energy conserving methods*, BIT 36 (1996b) 122–134. *[V.5]*

S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal. 36 (1999) 1549–1570. *[VIII.2], [IX.5], [IX.7]*

J.R. Rice, *Split Runge–Kutta method for simultaneous equations*, J. Res. Nat. Bur. Standards 64B (1960) 151–170. *[VIII.4]*

H. Rubin & P. Ungar, *Motion under a strong constraining force*, Comm. Pure Appl. Math. 10 (1957) 65–87. *[XIV.3]*

C. Runge, *Ueber die numerische Auflösung von Differentialgleichungen*, Math. Ann. 46 (1895) 167–178. *[II.1]*

H. Rüssmann, *On optimal estimates for the solutions of linear partial differential equations of first order with constant coefficients on the torus*, Dyn. Syst., Theor. Appl., Battelle Seattle 1974 Renc., Lect. Notes Phys. 38 (1975) 598–624. *[X.4]*

H. Rüssmann, *On optimal estimates for the solutions of linear difference equations on the circle*, Celest. Mech. 14 (1976) 33–37. *[X.4]*

R.D. Ruth, *A canonical integration technique*, IEEE Trans. Nuclear Science NS-30 (1983) 2669–2671. *[II.5], [VI.1], [VI.3], [IX.1]*

J.-P. Ryckaert, G. Ciccotti & H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*, J. Comput. Phys. 23 (1977) 327–341. *[VII.1], [XIII.1]*

P. Saha & S. Tremaine, *Symplectic integrators for solar system dynamics*, Astron. J. 104 (1992) 1633–1640. *[V.3]*

S. Saito, H. Sugiura & T. Mitsui, *Butcher's simplifying assumption for symplectic integrators*, BIT 32 (1992) 345–349. *[IV.10]*

J. Sand, *Methods for starting iteration schemes for implicit Runge–Kutta formulae*, Computational ordinary differential equations (London, 1989), Inst. Math. Appl. Conf. Ser. New Ser., 39, Oxford Univ. Press, New York, 1992, 115–126. *[VIII.6]*

J.M. Sanz-Serna, *Runge–Kutta schemes for Hamiltonian systems*, BIT 28 (1988) 877–883. *[VI.4]*

J.M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: an overview*, Acta Numerica 1 (1992) 243–286. *[IX.1]*

J.M. Sanz-Serna, *An unconventional symplectic integrator of W. Kahan*, Appl. Numer. Math. 16 (1994) 245–250. *[VII.4]*

J.M. Sanz-Serna & L. Abia, *Order conditions for canonical Runge–Kutta schemes*, SIAM J. Numer. Anal. 28 (1991) 1081–1096. *[IV.10]*

J.M. Sanz-Serna & M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, London, 1994. *[VI.3], [VIII.6]*

R. Scherer, *A note on Radau and Lobatto formulae for O.D.E:s*, BIT 17 (1977) 235–238. *[II.3]*

T. Schlick, *Some failures and successes of long-timestep approaches to biomolecular simulations*, in Computational Molecular Dynamics: Challenges, Methods, Ideas (P. Deuflhard et al., eds.), Springer, Berlin 1999, 227–262. *[XIII.1]*

M.B. Sevryuk, *Reversible systems*, Lecture Notes in Mathematics, 1211. Springer-Verlag, 1986. *[XI.0]*

L.F. Shampine, *Conservation laws and the numerical solution of ODEs*, Comp. Maths. Appls. 12B (1986) 1287–1296. *[IV.1]*

Z. Shang, *Generating functions for volume-preserving mappings and Hamilton–Jacobi equations for source-free dynamical systems*, Sci. China Ser. A 37 (1994a) 1172–1188. *[VI.9]*

Z. Shang, *Construction of volume-preserving difference schemes for source-free systems via generating functions*, J. Comput. Math. 12 (1994b) 265–272. *[VI.9]*

Z. Shang, *KAM theorem of symplectic algorithms for Hamiltonian systems*, Numer. Math. 83 (1999) 477–496. *[X.6]*

Z. Shang, *Resonant and Diophantine step sizes in computing invariant tori of Hamiltonian systems*, Nonlinearity 13 (2000) 299–308. *[X.6]*

Q. Sheng, *Solving linear partial differential equations by exponential splitting*, IMA J. Numer. Anal. 9 (1989) 199–212. *[III.3]*

C.L. Siegel & J.K. Moser, *Lectures on Celestial Mechanics*, Grundlehren d. math. Wiss. vol. 187, Springer-Verlag 1971; First German edition: C.L. Siegel, *Vorlesungen über Himmelsmechanik*, Grundlehren vol. 85, Springer-Verlag, 1956. *[VI.1], [VI.5], [VI.6]*

J.C. Simo & N. Tarnow, *The discrete energy-momentum method. Conserving algorithms for nonlinear elastodynamics*, Z. Angew. Math. Phys. 43 (1992) 757–792. *[V.5]*

J.C. Simo, N. Tarnow & K.K. Wong, *Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics*, Comput. Methods Appl. Mech. Eng. 100 (1992) 63–116. *[V.5]*

H.D. Simon & H. Zha, *Low rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput. 21 (2000) 2257–2274. *[IV.9]*

R.D. Skeel & C.W. Gear, *Does variable step size ruin a symplectic integrator?*, Physica D60 (1992) 311–313. *[VIII.2]*

M. Sofroniou & G. Spaletta, *Derivation of symmetric composition constants for symmetric integrators*, J. of Optimization Methods and Software (2004) to appear. *[V.3]*

A. Sommerfeld, *Mechanics* (Lectures on Theoretical Physics, vol. I), first German ed. 1942, English transl. by M.O. Stern, Acad. Press. *[VII.5]*

S. Sternberg, *Celestial Mechanics*, Benjamin, New York, 1969. *[X.0]*

E. Stiefel, *Richtungsfelder und Fernparallelismus in n-dimensionalen Mannigfaltigkeiten*, Comment. Math. Helv. 8 (1935) 305–353. *[IV.9]*

H.J. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin, 1973. *[II.3], [II.4], [V.1], [V.2]*

D. Stoffer, *On reversible and canonical integration methods*, SAM-Report No. 88-05, ETH-Zürich, 1988. *[V.1]*

D. Stoffer, *Variable steps for reversible integration methods*, Computing 55 (1995) 1–22. *[VIII.2], [VIII.3]*

D. Stoffer, *General linear methods: connection to one step methods and invariant curves*, Numer. Math. 64 (1993) 395–407. *[XV.2]*

D. Stoffer, *On the qualitative behaviour of symplectic integrators. III: Perturbed integrable systems*, J. Math. Anal. Appl. 217 (1998) 521–545. *[XII.4]*

C. Störmer, *Sur les trajectoires des corpuscules électrisés*, Arch. sci. phys. nat., Genève, vol. 24 (1907) 5–18, 113–158, 221–247. *[I.1]*

G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. 5 (1968) 506–517. *[II.5]*

W.B. Streett, D.J. Tildesley & G. Saville, *Multiple time step methods in molecular dynamics*, Mol. Phys. 35 (1978) 639–648. *[VIII.4]*

A.M. Stuart & A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1996. *[XII.3]*

G. Sun, *Construction of high order symplectic Runge–Kutta Methods*, J. Comput. Math. 11 (1993a) 250–260. *[IV.2]*

G. Sun, *Symplectic partitioned Runge–Kutta methods*, J. Comput. Math. 11 (1993b) 365–372. *[II.2], [IV.2]*

G. Sun, *A simple way constructing symplectic Runge–Kutta methods*, J. Comput. Math. 18 (2000) 61–68. *[VI.10]*

K.F. Sundman, *Mémoire sur le problème des trois corps*, Acta Math. 36 (1912) 105–179. *[VIII.2]*

Y.B. Suris, *On the conservation of the symplectic structure in the numerical solution of Hamiltonian systems* (in Russian), In: Numerical Solution of Ordinary Differential Equations, ed. S.S. Filippov, Keldysh Institute of Applied Mathematics, USSR Academy of Sciences, Moscow, 1988, 148–160. *[VI.4]*

Y.B. Suris, *The canonicity of mappings generated by Runge–Kutta type methods when integrating the systems $\ddot{x} = -\partial U/\partial x$*, Zh. Vychisl. Mat. i Mat. Fiz. 29, 202–211 (in Russian); same as U.S.S.R. Comput. Maths. Phys. 29 (1989) 138–144. *[VI.4]*

Y.B. Suris, *Hamiltonian methods of Runge–Kutta type and their variational interpretation* (in Russian), Math. Model. 2 (1990) 78–87. *[VI.6]*

Y.B. Suris, *Partitioned Runge–Kutta methods as phase volume preserving integrators*, Phys. Lett. A 220 (1996) 63–69. *[VI.9]*

Y.B. Suris, *Integrable discretizations for lattice systems: local equations of motion and their Hamiltonian properties*, Rev. Math. Phys. 11 (1999) 727–822. *[VII.2]*

Y.B. Suris, *The Problem of Integrable Discretization: Hamiltonian Approach*, Progress in Mathematics 219, Birkhäuser, Basel, 2003. *[X.3]*

G.J. Sussman & J. Wisdom, *Chaotic evolution of the solar system*, Science 257 (1992) 56–62. *[I.2]*

M. Suzuki, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A 146 (1990) 319–323. *[II.4], [II.5]*

M. Suzuki, *General theory of fractal path integrals with applications to many-body theories and statistical physics*, J. Math. Phys. 32 (1991) 400–407. *[III.3]*

M. Suzuki, *General theory of higher-order decomposition of exponential operators and symplectic integrators*, Phys. Lett. A 165 (1992) 387–395. *[II.5], [V.6]*

M. Suzuki, *Quantum Monte Carlo methods and general decomposition theory of exponential operators and symplectic integrators*, Physica A 205 (1994) 65–79. *[V.3]*

M. Suzuki & K. Umeno, *Higher-order decomposition theory of exponential operators and its applications to QMC and nonlinear dynamics*, In: Computer Simulation Studies in Condensed-Matter Physics VI, Landau, Mon, Schüttler (eds.), Springer Proceedings in Physics 76 (1993) 74–86. *[V.3]*

W.W. Symes, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, Physica D 4 (1982) 275–280. *[IV.3]*

F. Takens, *Motion under the influence of a strong constraining force*, Global theory of dynamical systems, Proc. int. Conf., Evanston/Ill. 1979, Springer LNM 819 (1980) 425–445. *[XIV.3]*

Y.-F. Tang, *The symplecticity of multi-step methods*, Computers Math. Applic. 25 (1993) 83–90. *[XV.4]*

Y.-F. Tang, *Formal energy of a symplectic scheme for Hamiltonian systems and its applications (I)*, Computers Math. Applic. 27 (1994) 31–39. *[IX.3]*

Y.-F. Tang, V.M. Pérez-García & L. Vázquez, *Symplectic methods for the Ablowitz–Ladik model*, Appl. Math. Comput. 82 (1997) 17–38. *[VII.4]*

B. Thaller, *Visual Quantum Mechanics. Selected topics with computer-generated animations of quantum-mechanical phenomena.* Springer-TELOS, New York, 2000. *[VII.6]*

W. Thirring, *Lehrbuch der Mathematischen Physik 1*, Springer-Verlag, 1977. *[X.5]*

M. Toda, *Waves in nonlinear lattice*, Progr. Theor. Phys. Suppl. 45 (1970) 174–200. *[X.1]*

J. Touma & J. Wisdom, *Lie–Poisson integrators for rigid body dynamics in the solar system*, Astron. J. 107 (1994) 1189–1202. *[VII.5]*

H.F. Trotter, *On the product of semi-groups of operators*, Proc. Am. Math. Soc.10 (1959) 545–551. *[II.5]*

M. Tuckerman, B.J. Berne & G.J. Martyna, *Reversible multiple time scale molecular dynamics*, J. Chem. Phys. 97 (1992) 1990–2001. *[VIII.4], [XIII.1]*

V.S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974 *[III.4], [IV.6], [IV.8]*

L. Verlet, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules*, Physical Review 159 (1967) 98–103. *[I.1], [XIII.1]*

A.P. Veselov, *Integrable systems with discrete time, and difference operators*, Funktsional. Anal. i Prilozhen. 22 (1988) 1–13, 96; transl. in Funct. Anal. Appl. 22 (1988) 83–93. *[VI.6]*

A.P. Veselov, *Integrable maps*, Russ. Math. Surv. 46 (1991) 1–51. *[VI.6]*

R. de Vogelaere, *Methods of integration which preserve the contact transformation property of the Hamiltonian equations*, Report No. 4, Dept. Math., Univ. of Notre Dame, Notre Dame, Ind. (1956) *[I.1], [VI.3]*

V. Volterra, *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*, Mem. R. Comitato talassografico italiano, CXXXI, 1927; *Opere* 5, p. 1–111. *[I.1]*

J. Waldvogel & F. Spirig, *Chaotic motion in Hill's lunar problem*, In: A.E. Roy and B.A. Steves, eds., *From Newton to Chaos: Modern Techniques for Understanding and Coping with Chaos in N-Body Dynamical Systems* (NATO Adv. Sci. Inst. Ser. B Phys., 336, Plenum Press, New York, 1995). *[VIII.2]*

G. Wanner, *Runge–Kutta-methods with expansion in even powers of h*, Computing 11 (1973) 81–85. *[II.3], [V.2]*

R.A. Wehage & E.J. Haug, *Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems*, J. Mechanical Design 104 (1982) 247–255. *[IV.5]*

J.M. Wendlandt & J.E. Marsden, *Mechanical integrators derived from a discrete variational principle*, Physica D 106 (1997) 223–246. *[VI.6]*

H. Weyl, *The Classical Groups*, Princeton Univ. Press, Princeton, 1939. *[VI.2]*

H. Weyl, *The method of orthogonal projection in potential theory*, Duke Math. J. 7 (1940) 411–444. *[VI.9]*

J.H. Wilkinson, *Error analysis of floating-point computation*, Numer. Math. 2 (1960) 319–340. *[IX.0]*

J. Wisdom & M. Holman, *Symplectic maps for the N-body problem*, Astron. J. 102 (1991) 1528–1538. *[V.3]*

J. Wisdom, M. Holman & J. Touma, *Symplectic correctors*, in *Integration Algorithms and Classical Mechanics* 10, J.E. Marsden, G.W. Patrick & W.F. Shadwick, eds., Amer. Math. Soc., Providence, R.I. (1996) 217–244. *[V.3]*

K. Wright, *Some relationships between implicit Runge–Kutta, collocation and Lanczos τ methods, and their stability properties*, BIT 10 (1970) 217–227. *[II.1]*

K. Wright, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math. 63 (1992) 283–295. *[IV.9]*

W.Y. Yan, U. Helmke & J.B. Moore, *Global analysis of Oja's flow for neural networks*, IEEE Trans. Neural Netw. 5 (1994) 674–683. *[IV.9]*

H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A 150 (1990) 262–268. *[II.4], [II.5], [III.4], [III.5], [V.3]*

H. Yoshida, *Recent progress in the theory and application of symplectic integrators*, Celestial Mech. Dynam. Astronom. 56 (1993) 27–43. *[IX.1], [IX.4], [IX.8]*

A. Zanna, *Collocation and relaxed collocation for the Fer and the Magnus expansions*, SIAM J. Numer. Anal. 36 (1999) 1145–1182. *[IV.7], [IV.10]*

A. Zanna, K. Engø & H.Z. Munthe-Kaas, *Adjoint and selfadjoint Lie-group methods*, BIT 41 (2001) 395–421. *[V.4], [V.6]*

K. Zare & V. Szebehely, *Time transformations in the extended phase-space*, Celestial Mechanics 11 (1975) 469–482. *[VIII.2]*

C. Zener, *Non-adiabatic crossing of energy levels*, Proc. Royal Soc. London, Ser. A 137 (1932) 696–702. *[XIV.1]*

S.L. Ziglin, *The ABC-flow is not integrable for $A = B$*, Funktsional. Anal. i Prilozhen. 30 (1996) 80–81; transl. in Funct. Anal. Appl. 30 (1996) 137–138. *[VI.9]*

Index

ABC flow 228
Abel–Liouville–Jacobi–Ostrogradskii identity 105, 228
Ablowitz–Ladik model 273
action integral 205
action-angle variables 397
adaptive Verlet method 309
adiabatic integrator 547
adiabatic invariants 531, 533, 545, 562
adiabatic transformations 531, 532
adjoint method 42, 145, 342, 613
– of collocation method 146
– of Runge–Kutta method 147
– quadratic invariants 176
adjoint operator 83
angular momentum 9, 98, 100, 101, 276, 591, 601
area preservation 5, 6, 183, 184, 188
Argon crystal 19
Arnold–Liouville theorem 397
attractive invariant manifold 460, 574, 610
attractive invariant torus 464
– of numerical integrator 467
averaged forces 319
averaging
– basic scheme 458
– perturbation series 459
averaging principle 456
avoided crossing 535, 563

B-series 51, 56, 57, 212, 223, 575
– composition 61
– symplectic 217, 219
backward error analysis 337, 576
– formal 337
– rigorous 360
BCH formula 83, 84, 348
– symmetric 86
Bernoulli numbers 84, 122
bi-coloured trees 66
B_∞-series 72
Birkhoff normalization
– Hamiltonian 412
– reversible 447
$B(p)$ 32
Butcher group 64, 372
Butcher product 75, 212

canonical 186
– equations of motion 181
– form 267
– Poisson structure 254
– transformation 186
canonical coordinates of a Lie group
– first kind 129
– second kind 129
Casimir function 257, 267, 283
Cayley transform 107, 128
central field 392, 400, 438, 440
characteristic lines 262
Choleski decomposition 154
coadjoint orbit 287
collocation methods 27, 30, 122
– discontinuous 35, 247
– symmetric 146
collocation polynomial 30
commutator 118
– matrix 83
compensated summation 323
complete systems 263
completely integrable 393
composition
– of B-series 61
– of Runge–Kutta methods 59
composition methods 43, 45, 50, 92, 105, 190, 333
– ρ-compatibility 145
– local error 150
– of order 2 150
– of order 4 152, 155
– of order 6 153, 156
– of order 8 157

– of order 10 158
– order conditions 71, 75, 80
– symmetric 149
– symmetric-symmetric 154
– with symmetric method 154
conditionally periodic flow 399
configuration manifold 239
conjugate momenta 181
conjugate symplecticity 222, 225, 592
conservation
– of area 5, 183
– of energy 98, 172, 366, 484, 512, 600
– of linear invariants 99
– of mass 98
– of momentum 172, 600
– of quadratic invariants 101, 102, 212, 214, 216
– of volume 227
conserved quantity 97
consistent initial values 238
constant direction of projection 165
constrained Hamiltonian systems 239, 258
constrained mechanical systems 237
continuous output 326
coordinates
– generalized 180
cotangent bundle 240
$C(q)$ 32
Crouch-Grossman methods 124
– order conditions 124

d'Alembert principle 259
Darboux–Lie theorem 261, 265, 266, 272
degrees of freedom 5
diagonally implicit Runge–Kutta methods
– symmetric 147
differential equations 2
– Hamilton–Jacobi 200
– Hamiltonian 4, 180
– highly oscillatory 21
– modified 337
– on Lie groups 118
– on manifolds 115, 239
– partial, linear 262
– reversible 143
– second order 7, 41, 216, 332
differential equations on manifolds
– ρ-compatibility 145
differential form 186
differential-algebraic equations 140, 237
diophantine frequencies 406
Dirac–Frenkel variational principle 138, 259, 295
DIRK methods
– symmetric 147
discontinuous collocation 35, 247
discrete Euler–Lagrange equations 206
discrete Lagrangian 206
discrete momenta 206
discrete-gradient methods 171, 174
dissipative systems 455
distinguished functions 266
divergence-free vector fields 227

eccentricity 9
effective order
– of composition methods 158
EI 150
elementary differentials 52, 53, 66
elementary Hamiltonian 373, 384
elementary weights 55
energy
– oscillatory 479, 484, 505, 510, 517, 524
– total 182, 479, 484, 510, 524
energy conservation 366, 379, 510, 524, 600
energy exchange 483, 490, 494
energy-momentum methods 171
– for N-body systems 173
equistage approximation 329
error analysis
– backward 337
error growth
– linear 413, 414, 448
– of rounding errors 324
Euler equations 275, 277, 279
Euler method
– –Lie 126
– explicit 3
– implicit 3
– symplectic 4, 48, 189, 230, 242, 270
Euler parameters 281
Euler–Lagrange equations 181, 205, 237
– discrete 206
explicit symmetric methods 148
exponential map 120

Fermi–Pasta–Ulam problem 21, 479
filter function 481
first integrals 5, 97, 211, 375
– long-time near-preservation 413, 448
– quadratic 212, 591
fixed-point iteration 330
flow 2
– discrete 3
– exact 2, 49, 200

– isospectral 107
– numerical 3, 49
– Poisson 261, 265
frequencies 399
– diophantine 406
Frobenius norm 132

G-symplectic 587
Gauss methods 34, 101, 333
– symmetric 147
– symplectic 192
Gaussian wavepacket 296
Gautschi-type methods 473, 477
general linear methods 609
– strictly stable 609
– symmetric 611
– weakly stable 610
generalized coordinate partitioning 117
generating functions 195, 197, 201, 204, 288, 344
– for partitioned RK methods 199
– for Runge–Kutta methods 198
geometrical numerical algebra 131
GL(n), general linear group 119
$\mathfrak{gl}(n)$, Lie algebra of $n \times n$ matrices 119
Grassmann manifold 131, 135
growth parameter 592, 614

Hénon–Heiles problem 380
Hall set 78
Hamilton's principle 204, 205
Hamilton–Jacobi equation 200, 391
Hamiltonian 4, 181, 257
– elementary 373, 384
– global 186
– local 185, 234
– modified 343, 375
Hamiltonian perturbation theory 389, 404
– basic scheme 405
– Birkhoff normalization 412
– KAM theory 410, 423
– perturbation series 406
Hamiltonian systems 4, 180
– constrained 239, 258, 289
– integrable 390
– non-canonical 237
– perturbed integrable 404
harmonic oscillator
– varying frequency 546
heavy top 283
Hénon–Heiles model 15
Hopf algebra 65

IE 150

implementation 303, 325
implicit midpoint rule 3, 34, 190, 192, 223, 270
– averaged 537
– symmetry 145
– symplecticity 190
impulse method 317, 475, 550
– mollified 476
index reduction 239, 241
inertia ellipsoid 275
integrability lemma 186
integrable systems 601
– Hamiltonian 390
– reversible 437
invariant manifold 574
– attractive 460, 574, 610
invariant torus 397, 423
– long-time near-preservation 422, 451
– of numerical integrator 433, 453, 467
– of reversible map 451
– of symplectic map 431
– weakly attractive 464
invariants 2, 5, 97
– adiabatic 531, 533, 545, 562
– linear 99
– polynomial 105
– quadratic 101
– weak 109
involution
– first integrals in 391
irreducible
– Runge–Kutta methods 220
isospectral flow 107, 403
isospectral methods 107
iteration
– fixed-point 330
– Newton-type 331

Jacobi identity 118, 255

KAM theory
– Hamiltonian 410, 423
– reversible 445
– reversible near-identity map 451
– symplectic near-identity map 431
KAM torus
– sticky 412
Kepler problem 8, 25, 46, 111, 150, 234, 416, 603
– perturbed 12, 26, 304
Kepler's second law 9
kernel
– of processing methods 158

kinetic energy 180, 237
Kolmogorov's iteration 410
Kolmogorov's theorem 423

Lagrange equations 181
Lagrange multipliers 111, 132, 237, 279
Lax pair 403
leap-frog method 7
left-invariant 289
Legendre transform 181
– discrete 206
Leibniz' rule 255
Lennard–Jones potential 19
Lie algebras 118, 286
Lie bracket 89, 118, 261
– differential operators 89
Lie derivative 87, 348, 362
– of B-series 370
– of P-series 382
Lie group methods 123, 351
– symmetric 169
Lie groups 118
– quadratic 128
Lie midpoint rule 127
Lie operator 261
Lie–Euler method 126
Lie–Poisson reduction 289
Lie–Poisson systems 274, 286
Lie–Trotter splitting 47
Lindstedt–Poincaré series 406
linear error growth 12, 413, 414, 448, 601
linear multistep methods
– weakly stable 575
linear stability 23
Liouville lemma 392
Liouville's theorem 227
Lobatto IIIA - IIIB pair 102, 192, 210, 247, 352, 386
Lobatto IIIA methods 34, 377
– symmetric 147
Lobatto IIIA–IIIB pair 40
Lobatto IIIB methods 37, 377, 449
– symmetric 147
Lobatto IIIS 235
Lobatto quadrature 247
local coordinates 113
– existence of numerical solution 167
– symmetric methods 166
local error 29
– of composition methods 150, 176
long-time behaviour
– symmetric integrators 437, 455
– symplectic integrators 389, 455
long-time energy conservation 366
Lorenz problem 176
Lotka–Volterra problem 1, 24, 175, 257, 270, 271, 273, 340
low-rank approximation 137
Lyapunov exponents 131

Magnus series 121
manifold of rank k matrices 131
manifolds 109, 114, 239, 267
– symmetric methods 161
– symplectic 258
Marchuk splitting 47
matrix commutator 83
matrix exponential 120
matrix Lie groups 118
mechanical systems 555
– constrained 237, 258
merging product 75
methods based on local coordinates 166
methods on manifolds 97, 350
– symmetric 161
midpoint rule 123
– explicit 569, 580
– implicit 3, 34, 190, 192, 223, 270
– Lie 127
– modified 171
modified differential equation 337
– B-series 369
– constrained Hamiltonian system 352
– first integrals 351
– Lie group methods 351
– Lie–Poisson integrators 354
– methods on manifolds 350
– P-series 381
– perturbed differential equation 466
– Poisson integrators 347
– reversible methods 343
– splitting methods 348
– symmetric methods 342
– symplectic methods 343
– trees 369
– variable steps 356
modified equation
– parasitic 579
modified Hamiltonian 343, 375, 589
– global 344, 353
modified midpoint rule 171
modulated Fourier expansion 496
– exact solution 486, 496
– Hamiltonian 503
– multi-frequency 519
– numerical solution 488, 498

molecular dynamics 18
mollified impulse method 476, 554
momenta 181
– conjugate 181
– discrete 206
moments of inertia 100
momentum
– angular 9, 98, 100, 101, 173
– linear 98, 173
momentum conservation 600
Moser–Veselov algorithm 281
multi-force methods 478
multi-value methods 609
– symmetric 611
multiple time scales 472, 479
multiple time stepping 316, 475
multirate methods 316
multistep methods 567
– backward error analysis 576
– G-symplectic 587
– partitioned 572
– second order equations 569
– strictly stable 568, 573
– symmetric 568, 570
– symplectic 585
– variable step sizes 605
Munthe-Kaas methods 125

N-body system 13, 98
– energy-momentum methods 173
Newton-type iteration 331
Noether's theorem 210
non-resonant frequencies 406
non-resonant step size 433, 498, 511
Nyström methods 41, 69, 96, 104
– symplectic 194

$\mathrm{O}(n)$, orthogonal group 119
one-leg methods 587
one-step method 8, 29, 187
– underlying 573, 609
optimal control 235
order 29
– of a tree 53, 67
– of symmetric local coordinates 167
– of symmetric projection 162
order conditions
– composition methods 71, 75, 80, 93, 94
– Crouch-Grossman methods 124
– Nyström methods 69
– partitioned RK methods 39, 69
– processing methods 159
– RK methods 29, 51, 56, 58
– splitting methods 80, 92
– symmetric composition 155
– symmetrized 177
ordered subtrees 60
ordered trees 60
oriented area 183
oriented free trees 388
orthogonal matrices 118
orthogonality constraints 131
oscillatory differential equations 21, 471, 531
oscillatory energy 22, 479, 484, 505, 510, 517, 524
outer solar system 8, 13, 112

P-series 68, 214
– symplectic 217, 219
parametrization
– tangent space 117
partial differential equations
– linear 262
partitioned Runge–Kutta methods 38, 102, 148
– diagonally implicit 149
– symmetric 148
– symplectic 193, 208, 231
partitioned systems 3, 66
pendulum 4, 5, 110, 181, 185, 188, 367, 396, 593
– double 233
– spherical 238, 254
– stiff spring 526
perturbation series
– averaging 459
– Hamiltonian 406
– reversible 444
perturbation theory
– dissipative 455
– Hamiltonian 389, 404
– reversible 437
phase space 2
Poincaré cut 16
Poisson
– bracket 255, 257
– flow 261, 265
– integrators 270, 272, 300
– maps 268
– systems 254, 257, 297
Poisson structures 265
– canonical 254
– general 256
polar decomposition 134
polynomial invariants 105

potential energy 181, 237
precession 12, 26
processing
– of composition methods 158
– order conditions 159
projection
– symplectic 259
projection methods 109, 351
– standard 110
– Stiefel manifolds 133
– symmetric 161
– symmetric non-reversible 166
pseudo-inverse of a matrix 116
pseudo-symplectic methods 436

QR algorithm 108
QR decomposition 134
quadratic invariants 101
– near conservation 225
quadratic Lie groups 128
quantum dynamics 293
quasi-periodic flow 399
quaternions 281

r-RESPA method 318, 475
Radau methods 34
rank k matrix manifold 131
RATTLE 245, 280, 352, 388
resonance
– numerical 482, 485, 602
resonance module 517
reversibility 239, 311
– of symmetric local coordinates 168
– of symmetric projection 163
reversible maps 143, 144
reversible methods 343
reversible perturbation theory 437
– basic scheme 443
– Birkhoff normalization 447
– KAM theory 445
– perturbation series 444
reversible systems 143
– integrable 437
– perturbed integrable 442
reversible vector fields 144
ρ-compatibility condition 145
ρ-reversible 143
– maps 144
– vector field 143
Riccati equation 134
rigid body 99, 163, 274, 280, 288, 441, 449
– Hamiltonian theory 278

Rodrigues formula 141
rooted trees 53
rounding error 322
Runge–Kutta methods 27, 28, 101, 311, 325, 333
– ρ-compatibility 145
– additive 50
– adjoint method 147
– implicit 29
– irreducible 220
– partitioned 38, 148
– symmetric 146
– symplectic 191, 231
Runge–Lenz–Pauli vector 26

s-stable 594
Schrödinger equation 293
– nonlinear 273
semiclassical dynamics 293
separable partitioned systems 231
SHAKE 245
simplifying assumptions 96
sinc function 473, 481
singular value decomposition 133
SL(n), special linear group 119, 130
$\mathfrak{sl}(n)$, special linear Lie algebra 119
small denominators 406
SO(n), special orthogonal group 119
$\mathfrak{so}(n)$, skew-symmetric matrices 119
spherical pendulum 238, 254
splitting
– fast-slow 317
– Lie–Trotter 47
– Marchuk 47
– of ordered tree 370
– Strang 47, 230
splitting methods 47, 48, 91, 193, 252, 270, 284, 298, 348
– ρ-compatibility 145
– negative steps 82
– of higher order 82
– order conditions 80
Sp(n), symplectic group 119
$\mathfrak{sp}(n)$, symplectic Lie algebra 119
stability
– linear 23
– long-term 592
stability function 194
starting approximations 326
– order 327
step size control
– integrating, reversible 314, 357, 449, 538

– proportional, reversible 310, 313, 356, 449
– standard 303
– structure-preserving 310
step size function 308, 311
Stiefel manifold 131
Störmer–Verlet scheme 7, 9, 39, 48, 189, 270, 318, 349, 386, 472, 586
– as classical limit 300
– as composition method 148
– as Nyström method 41
– as processing method 159
– as splitting method 48
– as variational integrator 208
– energy conservation 368, 513
– linear error growth 414
– symmetry 42, 145
– symplecticity 48, 190
– variable step size 308, 309, 312, 313, 315
Strang splitting 47, 230, 315, 348
structure constants 286
submanifold 109
– symplectic 259
subtrees
– ordered 60
summation
– compensated 323
superconvergence 32, 37, 250
Suzuki's fractals 45, 46, 153
switching lemma 76
symmetric collocation methods 146, 176
symmetric composition 94
– of first order methods 150
– of symmetric methods 150, 154
symmetric composition methods 149
– of order 6 156
– of order 8 157
– of order 10 158
symmetric Lie group methods 169
symmetric methods 3, 42, 143, 144, 342, 612
– explicit 148
– symmetric composition 154
symmetric methods on manifolds 161
symmetric projection 161
– existence of numerical solution 162
– non-reversible 166
symmetric Runge–Kutta methods 146, 176
symmetric splitting method 177
symmetrized order conditions 177
symmetry 289, 311, 613
– of Gauss methods 147
– of Lobatto 147
– of symmetric local coordinates 168
symmetry coefficient 57, 67, 72
symplectic 183, 196, 241
– B-series 217
– maps 268
– P-series 217
– projection 259
– submanifold 258, 295
symplectic Euler method 4, 48, 189, 193, 230, 242, 270, 340, 346, 349, 383
– as splitting method 48
– energy conservation 368
– variable step size 307
symplectic methods 187, 612
– as variational integrators 207
– based on generating functions 203
– irreducible 222
– Nyström methods 194
– partitioned Runge–Kutta methods 193, 208
– Runge–Kutta methods 191
– variable step size 306
symplectic submanifold 259
symplecticity 244, 585

Takens chaos 563
tangent bundle 239
tangent space 114, 120
– parametrization 117, 134
θ-method 147
– adjoint 148
three-body problem 321, 390
time transformation 306, 356
time-reversible methods 144
Toda flow 109
Toda lattice 402, 414, 440, 449
total differential 186, 196
total energy 9, 18, 21, 98, 479, 484, 510, 524, 600
transformations
– adiabatic 531, 532
– averaging 458
– canonical 186
– reversibility preserving 438
– symplectic 182, 183, 196, 241
trapezoidal rule 28, 194, 223, 312
trees 51, 217, 369
– bi-coloured 66
– equivalence class 384
– ordered 60
– oriented free 388

– rooted 53
∞-trees 72
trigonometric methods 473
triple jump 44, 46, 153
true anomaly 9
two-body problem 9, 25
two-force methods 478

underlying one-step method 573, 609

Van der Pol's equation 455
variational integrators 204
variational problem 205, 237
variational splitting 271
vector fields 2
– divergence-free 227
– reversible 143, 144
Verlet method 7, 39, 48, 189, 270, 318, 472, 513
– adaptive 309
Verlet-I method 318, 475
volume preservation 105, 113, 227, 231
volume-preserving integrators 228

weak invariants 109
work-precision diagrams 150, 153, 156, 157, 334, 336, 482, 604, 605, 608
W-transformation 235